University College London

Towards an understanding of the molecular dynamic causes of resistance in the HIV-1 Integrase

by

Milo A. Bem

A thesis submitted in partial fulfilment for the
degree of Doctor of Philosophy
in the Department of Chemistry

I, Milo Bem confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

21st February 2016

"Remember that all models are wrong; the practical question is how wrong do they have to be to not be useful." ([1] p 74)

Abstract

This work endeavours to explain molecular dynamic causes of resistance in Human Immunodeficiency Virus (HIV) Integrase through simulation of Prototype Foamy Virus (PFV) Intergrase interaction with DNA in the presence of Raltegravir (RLT) for the purpose of efficient and automatic drug ranking of Integrase inhibitors.

Chapter One introduces the problem of Acquired Immunodeficiency Syndrome (AIDS) through both historical perspective and current situation by some crude analysis of geographical and social distribution. Similarly HIV is introduced both as a biological species and biochemical system with particular focus on Integrase, one of its essential protein enzymes.

Chapter Two contains methods of **Molecular Dynamics** which are used extensively throughout this thesis (Chapters Four to Six), as well as some scientific background to those methods.

Chapter Three describes efforts towards generating viable structure of HIV Integrase intasome for the purpose of Molecular Dynamic analyses. Methods of **homology modelling** are employed using known homologues of HIV Integrase as input data.

The remaining parts of this thesis are made up of three experimental chapters describing Molecular Dynamic simulations of several biochemical systems and their analyses.

Contents

List of Tables

List of Figures

Acronyms

AIDS Acquired Immunodeficiency Syndrome. 1, 10–12

ASV Avian Sarcoma Virus. 49

CCD Core catalytic domain. 19, 49, 51

CHMP EU Committee for Medicinal Products for Human Use. 21

CTD C-terminal domain. 20, 49

ELV Elvitegravir. 19, 23, 24, 52

FDA United States Food and Drug Administration. 21

GAFF General Amber Forcefield. 64

HIV Human Immunodeficiency Virus. 1, 10–12, 14–18, 20–22, 24, 41, 45, 48, 49

IBD Integrase Binding Domain. 51

IN Integrase. 14, 15, 21, 41, 48, 62

INSTI Integrase strand transfer inhibitor. 19, 62

LEDGF Lens epithelium-derived growth factor. 51

LTR long terminal repeat. 17, 19, 45

MD Molecular Dynamics. 26, 41, 62

MM Molecular Mechanics. 27

MM-GBSA molecular mechanics generalised Born surface accessibility. 36, 38, 62, 69

MM-PBSA molecular mechanics Poisson–Boltzmann surface accessibility. 36–38, 62, 69

MVV Maedi-Visna Virus. 51

NED N-terminal extension domain. 20

NMR Nuclear Magnetic Resonance. 26, 48

NTD N-terminal domain. 19, 20, 51

PCA principal components analysis. 69

PDB Protein Data Bank. 47

PFV Prototype Foamy Virus. 1, 52, 62

PIC pre-integration complex. 19

PR Protease. 15

QM Quantum Mechanics. 26

RESP Restrained Electrostatic Potential. 64

RLT Raltegravir. 1, 19, 21–24, 38, 52, 62, 63, 69

RMSD root mean square deviation. 34, 66, 67, 85

RMSF root mean square fluctuation. 34, 70

RSV Rous Sarcoma Virus. 18, 49

RT Reverse transcriptase. 15

SIV Simian Immunodeficiency Virus. 13, 47, 50

1. Introduction

1.1. Historical outline

The oldest —retrospectively identified— case of AIDS infected patients in Europe are a three-member Norwegian family, who all died from AIDS in 1970s. Although their routes of infection were not proven, the family history suggests the man contracted HIV during his visit in Africa as a sailor before 1966 and subsequently infected his wife, who as a result gave birth to an infected daughter[2]. Several other cases were identified in tissue samples collected from various African patients during 60s–70s[3].

AIDS was diagnosed and defined for first time in USA in 1981. The syndrome was initially known as **Gay-related immune deficiency** (GRID), or more colloquially as the *gay plague*, as it was mostly diagnosed within the gay communities of California and New York. It had been found that the main routes of contagion were homosexual intercourse[4], intravenous drug administration[5] and blood transfusion. The name was changed to AIDS upon growing number of diagnoses with heterosexual men and women. HIV was isolated from AIDS patients in 1983[6][7].

In 2008 alone approximately 2 million people died from AIDS related causes and an estimated 38.6 million were living with HIV around the world[8], the overwhelming majority of whom live in Sub-Saharan Africa[9]. The number of new infections peaked around 1997 and has kept falling ever since, thanks to education, prevention and introduction of antiretroviral therapies. Similarly, the number of AIDS-related

deaths is falling since 2004, thanks to both the decreasing number of new infections and more efficient therapies. In developed countries AIDS is beginning to be rather treated as a chronic than deadly disease. The total number of HIV infected individuals has probably peaked around 2009 within the measure of uncertainty[9].

1.2. Distribution and mortality

The total HIV adult prevalence ratio in the world is estimated at 0.8% but the distribution is not uniform. Above average ratios are found in Sub-Saharan Africa (5%) and the Caribbean (1%) and the lowest in East Asia (0.1%). Sub-Saharan Africa is the largest pool of HIV infections with more than 22 million people living with HIV, or 68% of the global number of infected. It is also by far the most affected region in terms of AIDS-related deaths — 72% of global number of AIDS-related deaths in 2009 occurred in that region. For more detailed data see Table 1.1.

Region	Carriers	Infections	Prevalence	Deaths
Sub-Saharan Africa	22.5 mln	1.8 mln	5.0	1.3 mln
Middle East and North Africa	460 000	75 000	0.2	24 000
South and South East Asia	4.1 mln	270 000	0.3	260 000
East Asia	770 000	82 000	0.1	36 000
Oceania	57 000	4 500	0.3	1 400
Latin America	1.4 mln	92 000	0.5	58 000
Carribean	240 000	17 000	1.0	12 000
Eastern Europe and Central Asia	1.4 mln	130 000	0.8	76 000
Western and Central Europe	820 000	31 000	0.2	8 500
North America	1.5 mln	70 000	0.5	26 000
Total	33.3 mln	2.6 mln	0.8	1.8 mln

Table 1.1. Regional HIV and AIDS statistics for 2009; Total number of HIV carriers, new infections, adult prevalence ratio and AIDS-related deaths[9]

The main identified risk groups are male homosexuals, female sex-workers and intravenous drug users. The detailed breakdown between risk groups

vary strongly among regions and countries. In Sub-Saharan Africa, where 68% of carriers live, the majority of people living with HIV are women. The numbers are similar for Caribbean, Eastern Europe and Central Asia, Middle East and North Africa, and Oceania, where 40–50% of carriers are women. The same regions also have higher than average prevalence of HIV among children, mostly as a result of mother-to-child transmissions, with the exception of Eastern Europe and Central Asia.

Conversely, in highly developed countries of North America, Western and Central Europe, and East Asia, the disease remains more prevalent within men having sex with men, with women constituting less than 30% of infected. Similarly, prevalence of HIV among children in Western and Central Europe, and North America remains below 0.5% (Table 1.2).

Region	Total	Women	%	Children	%
Sub-Saharan Africa	22.5 mln	12.1 mln	54	2.3 mln	10
Middle East and North Africa	460 000	210 000	46	21 000	4.6
South and South East Asia	4.1 mln	1.4 mln	34	150 000	3.7
East Asia	770 000	220 000	29	8 000	1.0
Oceania	57 000	25 000	44	3 100	5.4
Latin America	1.4 mln	490 000	35	36 000	2.6
Carribean	240 000	120 000	50	17 000	7.1
Eastern Europe and Central Asia	1.4 mln	690 000	49	18 000	1.3
Western and Central Europe	820 000	240 000	29	1 400	0.2
North America	1.5 mln	270 000	18	4 500	0.3
Total	33.3 mln	15.8 mln	47	2.6 mln	7.7

Table 1.2. Regional HIV and AIDS statistics for 2009; Women and Children living with HIV as a percentage of infected population[9]

1.3. HIV

HIV was isolated in 1983 from a single AIDS patient and proposed as the infectious agent[6]. It is classified as a member of **lentiviruses** family, due to very long incubation period. It is generally accepted that

HIV evolved from Simian Immunodeficiency Virus (SIV) naturally found in African primates. Phylogenetic studies indicate there are two subspecies — HIV-1, more aggressive and more closely related to SIV strains found in the great ape *Chimpanzee*; and HIV-2, mostly limited to West Africa and closely related to the *Sooty mangabeys* strain of SIV. The data suggests the HIV species is polyphyletic, meaning two subspecies are more closely related to respective SIV strains than to each other and form only two small subbranches of the SIV clade[10]. The implication of this finding is that the virus adapted to humans at least twice. More detailed analysis reveals at least two transmissions took place for each of the subspecies[11].

SIV itself is been estimated to prey on various primate species for thousands of years but the transmissions to humans are most likely to have occurred in first half of the twentieth or late ninetieth century. The results of phylogenetic analyses put the estimated date of transmission around the 1950s, which would be consistent with the earliest identified infections.

1.3.1. Classification

Virus classification —despite using similar naming convention as for cellular organisms taxonomy— does not follow the same rules, partially due to the debated status of viruses themselves as living organisms. There are therefore many unrelated classification systems depending on features taken into account. Most existing classifications include both virus properties (like type of nucleic acid, morphology, modes of replication) and the host classification.

Retroviruses are a family (*Retroviridae*) of viruses. Their characteristic feature is the process of **reverse transcription**, which —when discovered[12][13]— caused modification of the *Central Dogma of Molecular Biology*[14], only thirteen years after it was proposed by Francis Crick at a symposium held at University College London in 1957[15]. After

entering the host cell, one of the virus enzymes *Reverse transcriptase*, translates viral RNA into DNA[12][13] which is then integrated into the host genome[16] by Integrase (IN). The integrated virus or *provirus* is ready to be replicated by the host whenever the infected part of genome is activated for transcription and expression. The third of essential retrovirus enzymes is *Protease* reponsible for cleaving expressed polypeptides into functional proteins.

Lentivirus —from Latin *lentus* for *slow, inactive*— is a genus of *Retroviridae* characterised by slow incubation. Their specific trait of having two —usually identical— single strands of RNA instead of one as most viruses do makes them in some aspects behave like diploid eukaryotes. When two different parent genomes are present they may recombine in a manner similar to that of higher organisms. Lentiviruses also have the unique ability to infect non-cycling cells, unlike other retroviruses which —even after entering the cell— can only integrate its genome if the nuclear membrane is broken (e.g. during natural cell division)[17]. Both these traits together make *Lentiviruses* very useful tools as vectors[18] e.g. for gene therapy[19][20]. The same traits, unfortunately, make HIV the causative agent for one of the most notorious pandemics of late 20th century.

HIV phylogenetic relations within the whole family is presented in closer details in chapter 3.

1.3.2. Structure

A virion of a typical retrovirus is an entity about 100 nm of diameter. Like many viruses it consists of a glycoprotein envelope, the genome and several enzymatic proteins as seen in Figure 1.1.

The outermost layer of a virion, the so called envelope, consists of a lipid membrane with glycoproteins transmembrane TM (gp41[1]) and surface envelope SU (gp120). TM–SU complex shows similarities to other

[1] all protein masses in kDa are based on HIV example

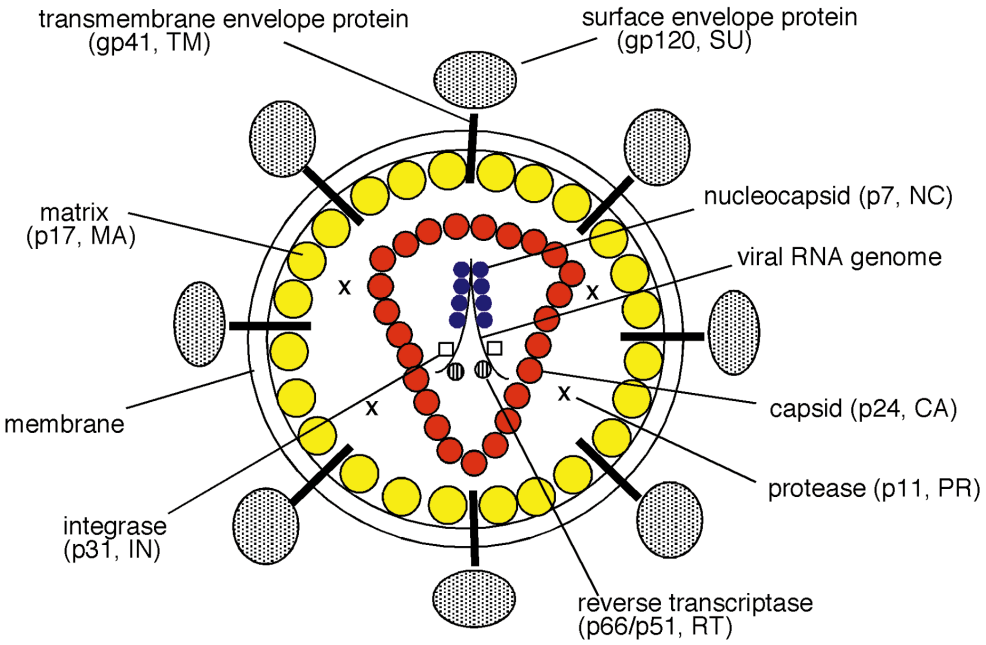

Figure 1.1. HIV-1 virion structure as a typical example of Lentivirus, adapted from[21].

viral membrane fusion proteins[22], with TM being a homotrimer build of three transmembrane α-helices and a functional external N-terminal fusion domain[23], and trimeric antigenic structure of SU[24]. The weak interaction within the complex is responsible for its short half-life of 30 hours and consequently rapid decay of virus infectivity[25].

In the virion one can find all three enzymes essential for its reproduction and maturation. Approximately 80 copies of those enzymes are found per virion[25]. Reverse transcriptase (RT), and IN are found packed closely with the RNA in the conical capsid. Protease (PR) is still active during maturation phase —after leaving the host— before the virion acquires its final shape.

The structure of a virion is shaped by three distinct proteins. **Matrix** protein MA (p17) forms icosahedral inner layer of the membrane[26]. A conical **Capsid** built of CA (p24) separates viral nucleus from lateral body of unknown composition[27]. Finally **Nucleocapsid** protein NC (p7) is packed closely with RNA and both enzymes essential for reproduction (RT, IN)[28][29].

Lentivirus genomes consist of two RNA strands, each coding the same three polypeptides **gag**, **pol**, **env** which are later processed into functional proteins, at different stages of the virus life cycle. There are also several smaller genes encoding helper proteins. The schematics of lentiviral genome at example of HIV is shown in Figure 1.2.

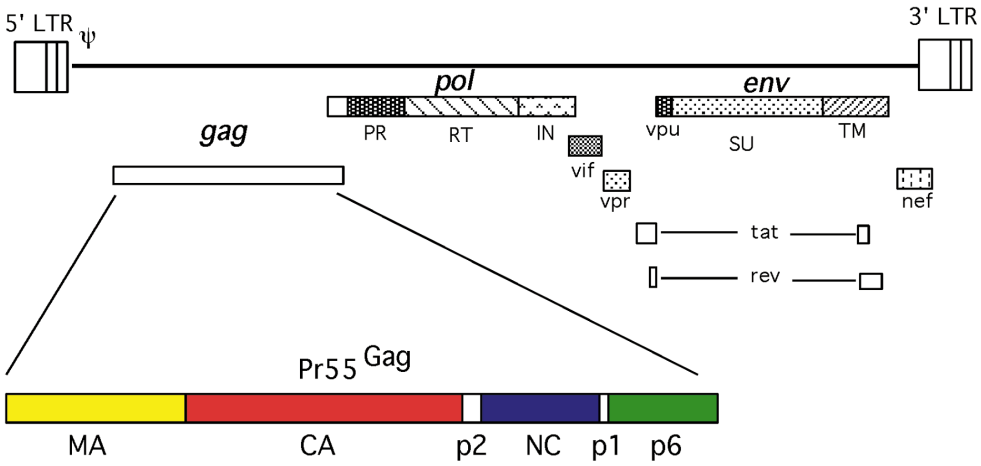

Figure 1.2. HIV-1 genome as a typical example of Lentivirus, adapted from[21].

The **gag** gene (*Group-specific Antigen*) encodes a precursor of structural capsid proteins, *NC*, *CA*, *MA*.

The **pol** gene (*Polyprotein*) encodes a precursor of viral enzymes, *Reverse transcriptase*, *Integrase* and *Protease*.

The **env** gene (*Envelope*) encodes a precursor of envelope proteins, *TM* and *SU*.

1.3.3. Lifecycle

HIV virion recognises and binds the CD4 receptor of the T4 lymphocyte[30]. The surface envelope protein responsible for interaction SU (gp120) is one of two proteins encoded by *env* gene. The high affinity between gp120 and CD4 causes downmodulation of CD4 receptors during later stage of infection and subsequently "superinfection interference"[31], although other factors have been suggested[32]. CD4 is not the only receptor identified as involved in virus entry[33]. The complex of gp120/CD4

and coreceptors is finally activated by TM (gp41, the other part of *env*) leading to membrane fusion[34][33]. It has been also shown that — at least HIV-2 — can enter cells stripped of CD4[35].

The action characteristic for retroviruses is performed by HIV RT, a heterodimer in which both parts (p66 and p51) are encoded by the same segment of *pol*. The shorter p51 is the result of cleaving out 15kDa C-term segment of p66 by PR. Because HIV, like all lentiviruses, has two strands of RNA, the RT, with its relatively low affinity to its template, is able to switch during the process of transcription[36]. This template switching combined with high mutation rate ($3 \cdot 10^{-5}$ per cycle[37]) together with their recombination, leads to high heterogeneity and subsequently to a rapid evasion of the host's natural immune response as well as of designed therapies[38].

The DNA is transported into cell nucleus as part of the pre-integration complex (PIC) but the mechanism of transfer, unique for Lentiviruses is not yet properly explained in detail[38]. After entering the nucleus HIV IN (p32) performs the process common among Retroviruses, of fusing viral DNA with that of the host. The integrated viral genome or *provirus* is then treated by the host as any other part of its own genome. The closer look at IN and the process will be presented in section 3.3.

Integrated viral DNA in the form of provirus is unrecognisable as alien for the host cell. It is indiscriminately expressed by the host. The viral long terminal repeat (LTR) (see Figure 1.1) act as an initiation sequence for transcription enzymes. The activity is significantly increased in the presence of Transcriptional transactivator (**tat**) protein[39][40]. Transcibed RNA is spliced in the process aided by the "regulator of expression of viral proteins" (**rev** protein)[41]. Transcription is followed by translation to viral protein precursors.

The assembly of new virion takes place at the host's membrane. Gag proteins are necessary for the successful assembly of virion-like particles

in vitro[21]. CA–NC complex *in vitro* selfassembly efficiency is significantly increased by the presence of RNA[42]. The *env* glycoprotein synthesised on ER[43][44] is subsequently oligomerised during its transport through the Golgi[45] and cleaved into TM and SU proteins[46][47]. Other interactions between viral proteins – essential for proper assembly – have been identified, too numerous to be presented in this work. Finally the assembled virion is released from the host in the process of budding[48].

The GagPol polyprotein processing (cleavage) by PR is the last stage of virus replication[49]. Unlike most other retroviruses, which mature after leaving the host[50], HIV protease probably performs its role before virion release from the host, during the final stage of assembly on the host membrane[51]. The maturation phase is reflected by large morphological change of the virion and emergence of a conical capsid around its genome[52].

1.4. Integrase

The importance of integration for lentiviruses was recognised since the discovery of Rous Sarcoma Virus (RSV) provirus identified in the DNA of infected cell homology to RSV genome[53] and confirmed by later research. It had been shown early that selective mutation in IN encoding segment of **pol** gene stops virus integration[54] and consequently its reproduction[55]. This fact makes Integrase potentially a good target for inhibition therapy.

1.4.1. Integration process

After transcription of viral RNA into DNA by RT the Integrase takes turn and the Integration begins. The process consists of three phases, first two of which are performed by Integrase[56].

3'-end processing

The integrase removes several nucleotides from 3'-end LTR of the viral DNA. This exposes the invariant CA_{OH}-3' dinucleotide that will participate in the integration. The preprocessed form of DNA bound to IN is called pre-integration complex (PIC).

Strand transfer

The stage begins with cleavage of cellular DNA by IN and completes with joining processed 3'-ends of viral DNA with cleaved 5'-end cellular DNA. This is the stage targeted by both available Integrase strand transfer inhibitors(INSTIs) — RLT and Elvitegravir (ELV)[57].

Repair

The final stage of integration is left to cellular enzymes. The host repair enzymes connect viral 5'-end with the open 3'-end of cellular DNA.

1.4.2. Domain organization

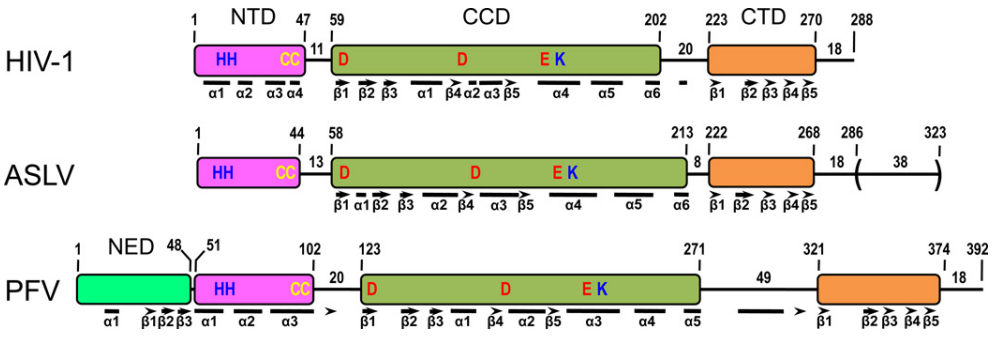

Figure 1.3. Domain organization and secondary structural elements of representative IN proteins, adapted from[57]. Domains color-coded and labelled. Also labelled HHCC residues of Zinc finger motif within NTD and DDE of catalytic triad in CCD.

The related retroviral integrases share common features of both sequence and structure[57]. The domain organization of selected members of the family presented in Figure 1.3 show three analogous domains — N-terminal domain (NTD) with Zinc finger motif, Core catalytic domain

(CCD) with catalytic DD(35)E and C-terminal domain (CTD). PFV differs from most integrases in having N-terminal extension domain (NED). In this work NED will be referred to as part of NTD for simplicity.

1.4.3. Common motifs

Zinc finger

Zinc finger motif (Znf) usually contains tetrahedral zinc cation increasing structural stability of a domain. In the class of Znf found in IN NTD the cation is chelated by two histidines (H) and two cysteines (C) and is thus called HHCC, Zn–HHCC, or Cys_2His_2.

D,D(35)E

Integrase's catalytic triad — D,D(35)E motif of two aspartates (D) and one glutamate (E). The number *35* indicates a conserved number of residues in between the second aspartate and glutamate, while the distance between two aspartates varies among species. In some resolved structures (described in chapter 3), one or two divalent cations have been found chelated by triad's residues (in which case also called Mg_2–DDE).

1.5. Therapies

The first approved treatment for HIV infected patients was Nucleotide Reverse Transcriptase Inhibitor (NRTI) **zidovudine** approved in 1987[58]. The drugs inhibit virus replication during the process of reverse transription, specific to retoviruses family. Shortly after, the first drug resistant HIV strains were reported[59]. This fact, combined with the high toxicity of early treatments led to a high concentration of efforts and resources on finding ever better therapy. In subsequent years, new types of therapies and new targets for drug therapies were acquired, Protease Inhibitors (PI) in 1995, Non-Nucleoside Reverse Transcriptase Inhibitors

(NNRTI) and Combination Therapies in 1997, Fusion Inhibitors in 2003 and finally Integrase Inhibitors in 2007.

The first Integrase inhibitor — RLT — was approved for use in USA market by United States Food and Drug Administration (FDA) in October 2007[60] and for the European market by the EU Committee for Medicinal Products for Human Use (CHMP) in December 2007. RLT inhibits the *strand transfer* stage of HIV DNA integration (section 1.4) by binding IN.

Regardless of drug category, the expected effect by slowing down the virus replication is reducing the number of virus infected cells in the patient's body, slowing the advance of disease and reducing the risk of further infections. When applied to pregnant and breastfeeding women, therapies are expected to reduce the risk of mother-to-child transmission[61].

1.5.1. Drug Design

The first use of medicine to cure or relief illnesses fades in the deep prehistory, hence no written account of the discovery. For ages the process of discovering cures was largely chaotic and done by untrained professionals or by pure chance. Similarly medical practitioners had close to no understanding of the nature of illnesses or drug workings.

Nowadays, drug design routinely uses atom level description of receptor structure as a target for a drug. The target usually being a molecule taking part in one of the critical pathways responsible for spreading or the advance of a disease. In the case of HIV, therapies typically target HIV enzymes, with the exception of *Fusion Inhibitors* or *CCR5 receptor antagonists* targeting process of virus fusion with the host cell. One of the first steps in efficient and accurate drug design relies heavily on detailed receptor structure to predict the interactions between the drug and the receptor. In case of lack of reliable structure, homologous protein or partial structure may be used as a first approximation.

1.5.2. Resistance

One of the biggest obstacles to eradication of viral diseases is the virus acquiring resistance to administered drugs. Viruses naturally show variation in genome, and occasionally if these mutations increase their resistance to the medicine present in the patient's system, the virus evolves toward more resistant to that particular drug by a means of natural selection[62]. HIV, being a Lentivirus, is a particularly evasive target due to a combination of reasons, among them having a doubled genome and long incubation period. Having a doubled genome gives lentiviruses not only a double chance of having resistance mutations but also a power of recombination in case of a patient infected with two strains of different origin. The causes for spreading of drug resistant HIV strains may be divided in two categories — patient noncompliance and nonsuppressive antiviral therapy[63].

Even during 48 week clinical BENCHMRK studies of RLT, two resistance pathways were observed in 105(23%) of 462 patients ([60], [64], [65]). The aim of this work is thus to use molecular simulations in drug binding characterisation, with the eventual goal of quantitatively describing the impact of mutations upon it.

Two main RLT resistant mutations are N155H and Q148H/R/K with 10- and 25-fold *in Vivo* resistance, respectively. Other mutations reported include L74M, E92Q, E138K, G140S/A, and G163R. For ELV, two main mutations are T66I and E92Q (37- and 36-fold reduced susceptibility, respectively). Other mutations reported are H51Y, T66I, Q95K, E138K, Q146P, S147G, and E157Q. Mutations selected from viral samples obtained from drug-resistant patients for the comparative study by Marinello *et al.* are shown in Figure 1.4 and their results showing great variance in efficacy in Figure 1.5.[66]

Growing number of drug resistant mutations led to increasing importance of *combination therapies* and is a motivating factor for development

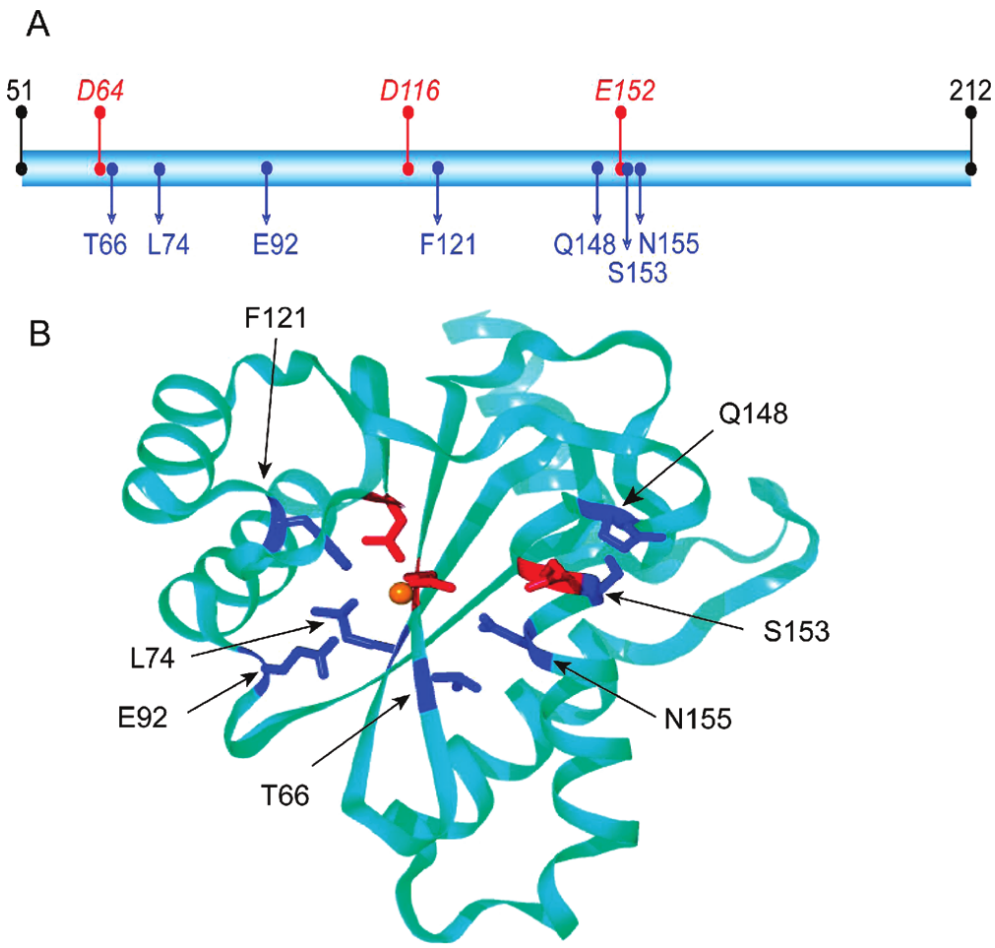

Figure 1.4. HIV-1 integrase CCD residues associated with resistance to RLT or ELV. (A) Linear representation. (B) Three-dimensional structure (PDB:1BI4). Catalytic acidic residues (DDE motif) are shown in red, mutations in dark blue. Figure adapted from [66].

of patient specific therapies. Since any mutation usually give resistance to a limited number of inhibitors —or to put it more accurately, a different level of resistance to different inhibitors— it is desired to choose a drug working most effectively towards the particular strain of virus. Virus genome is now routinely being sequenced upon a patient's diagnosis for consideration in clinical decisions. Growing knowledge databases and expert systems are aiding clinicians in decision making, but the rapid evolution of virus makes it important to develop more sophisticated, accurate and efficient methods. A number of projects exploring the possibility of modelling inhibitor binding to target receptor by means of **molecular dynamics** have been undertaken. It is postulated that

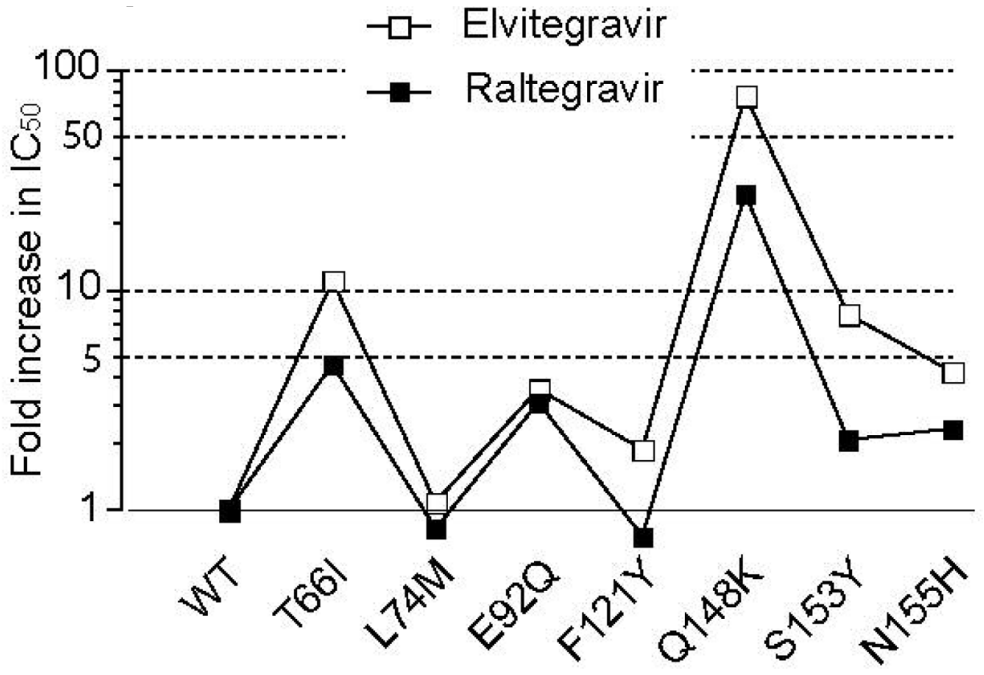

Figure 1.5. Comparative impact of the HIV-1 integrase mutations on integrase strand transfer inhibition by RLT and ELV. Figure adapted from [66].

inhibitor's adequacy toward a given mutant may be predicted by calculating its **binding free energy** with the receptor. The molecular dynamics methods, as well as methods for binding energy calculation are explored in chapter 2.

1.6. Goals

All known Integrase Inhibitors were successfully designed without a complete knowledge of the HIV Integrase structure. Several partial structures, as well as structures of homologous proteins are known. Having complete Integrase complex structure could have a tremendous positive efect on the drug design process, but the HIV Integrase heretofore eludes all eforts of crystallisation. It is therefore suggested to use theoretical models based on homologous proteins and partial HIV Integrase structures to research the behaviour of the drug target.

This thesis documents attempts to generate a viable homologous model for HIV Integrase and to provide a good explanation of bind-

ing active centre behaviour and structural factors behind drug resistant mutations by an analysis of the structural and thermodynamic properties of Prototype Foamy Virus engineered complex with DNA and Raltegravir, by using the methods of Molecular Dynamics and Free Binding Energy calculations.

2. Methods of Molecular Dynamics

The biochemical systems such as Integrase intasome are inherently dynamic. However, the information obtained through methods of X-ray crystallography and Nuclear Magnetic Resonance (NMR) is of a static nature. Therefore Molecular Dynamics (MD) methods are commonly employed in research to gain deeper insight in their dynamic structure and function through simulation.

Various approaches have been designed to fit the purpose of research of which only some will be explored here. MD methods can be used to refine structures by minimising systems' internal energy or to explain dynamics of chemical reactions.[67]

2.1. Quantum mechanics

Although Quantum Mechanics (QM) tells us that the *complete* description of any system is fundamentally beyond our cognition and classical mechanics is only its inaccurate approximation, it is rarely used in biochemistry due to very high computational cost. Evolution of quantum system is defined in terms of the *time dependent Schrödinger equation*:

$$i\hbar\frac{\partial\Psi}{\partial t} = \hat{H}\Psi \tag{2.1}$$

where Ψ is the wave function describing probability amplitude, $i\hbar\frac{\partial}{\partial t}$ is the *energy operator* and $\hat{H}$ is the *Hamiltonian operator* describing the system's energy.

To properly describe system evolution, one needs to find the wavefunction by solving the equation. The equation can only be solved analyt-

ically for the very simplest systems (e.g. a single particle in a potential, hydrogen atom). For a single particle in a potential the equation takes the form:

$$i\hbar\frac{\partial}{\partial t}\Psi(\mathbf{x},t) = -\frac{\hbar^2}{2m}\nabla^2\Psi(\mathbf{x},t) + \hat{V}(\mathbf{x})\Psi(\mathbf{x},t) \tag{2.2a}$$

$$\hat{H} = \hat{T} + \hat{V} \tag{2.2b}$$

$$\hat{T} = -\frac{\hbar^2}{2m}\nabla^2 \tag{2.2c}$$

$$\hat{V} = \text{potential energy operator} \tag{2.2d}$$

where $\mathbf{x}$ is particle position and t is time.

For bigger systems complexity grows exponentially, quickly making it unfeasible to solve the equation by analytical methods, and very expensive to compute numerically. Heuristic numerical methods using various approximations are therefore employed for medium-sized systems, including small organic molecules like drugs. However, most of the systems biochemistry regularly deals with are much more complex, with thousands of atoms involved in various interactions. Their quantum description cannot be hoped to be solved in reasonable time even using the most powerful computers.

2.2. Newtonian approach

Classical MD methods — also known as Molecular Mechanics (MM) — use simple Newtonian equations as reasonable approximations of mathematically advanced and resource-consuming quantum mechanics for the description of the behaviour of biomolecules. All atoms are described as points with mass, charge, position and velocity which can then be treated according to the classical laws of mechanics. Atom properties, positions and charges are used to compute their current potential and subsequently forces felt by each atom over the course of simulation. Forces acting on

atoms cause change of their momenta which, when integrated, are used to update positions and velocities at every step.

2.2.1. Equations of motion

The total mechanical energy of a system can be described by Hamiltonian equation as being a sum of the kinetic T_i and potential V_i components for all of the particles involved:

$$H = \sum_i \left(T_i(\dot{\mathbf{x}}) + V_i(\mathbf{x})\right) \tag{2.3a}$$

$$T_i = \frac{1}{2} m_i \dot{\mathbf{x}}_i^2 \tag{2.3b}$$

$$V_i = V_{bonded}(\mathbf{x}_i, \mathbf{x}_{j\ldots}) + V_{nonbonded}(\mathbf{x}_i, \mathbf{x}_{j\ldots}) + \ldots \tag{2.3c}$$

T_i is independent function of momentum for each particle and can be computed directly. Potential energy (V_i) however, being in principle a function of all the particles positions is usually only calculated in approximation (see section 2.4). For closed systems total energy is conserved, hence by comparing its time derivative to zero, one can derive an equation of system evolution as a form of Newton's second law of motion:

$$\frac{\partial H_i}{\partial t} = 0 \tag{2.4a}$$

$$\frac{\partial T_i(\dot{\mathbf{x}})}{\partial t} = -\frac{\partial V_i(\mathbf{x})}{\partial t} \tag{2.4b}$$

$$m\ddot{\mathbf{x}}_i = -\nabla V_i(\mathbf{x}_i, \mathbf{x}_{j\ldots}) \tag{2.4c}$$

The description of particles' potential energies is, therefore, all that is theoretically needed to calculate all of a system's past and future history given its state (positions and velocities) at any single moment. In practice, while positions are usually obtained from crystallographic structure or other similar methods, initial velocities are often randomised to simulate internal thermal motions.

2.2.2. Calculating forces

The potential component of the energy state of a molecular system is described by a *force field*, which is a bit misleading name for what is actually a conservative *potential field* (scalar field), of which the physical force field is a gradient (vector field). The actual forces acting on atoms are calculated during the course of the simulation using the force field equation, which in its simplest form is a sum of several interactions components, sometimes grouped into bonding and nonbonding for clarity:

$$V_{total} = V_{bonded} + V_{nonbonded} \tag{2.5a}$$

$$V_{bonded} = V_{bond} + V_{angle} + V_{dihedral} + \dots \tag{2.5b}$$

$$V_{nonbonded} = V_{Coulomb} + V_{vdW} + \dots \tag{2.5c}$$

The bonding components are those describing interactions within groups of atoms connected by covalent chemical bonds(Figure 2.1), while nonbonding are long-distance interactions (typically electrostatic and Van der Waals) between any atoms within a model.

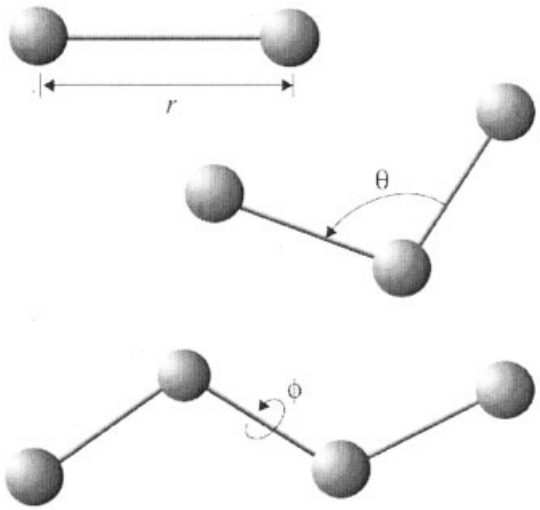

Figure 2.1. Simplified scheme of interactions within molecule. Generalised coordinates labelled: r - the bond length, θ - angle between subsequent bonds, and ϕ - dihedral angle between conjugated bonds

Actual values of energies for each of the parts of the equation are calculated using predefined set of parameters also called *Force field*, often with their own additional non standard terms. Generalised coordinates in the equations are calculated at each step from actual atom positions. Finally, all coordinates are used to compute forces according to the force field equation. The three most commonly used force fields in biochemisty are Amber[68], Gromos[69] and Charmm[70].

2.2.3. Integration algorithm

Because analytical solutions of system evolution are practically impossible for any biochemical system, updating of positions and velocities is done numerically in discrete steps instead. Given the current position, velocity and acceleration computed from potential through equation of motion, the next step of simulation can be approximated by Taylor expansion:

$$\mathbf{x}(t+\delta t) = \mathbf{x}(t) + \dot{\mathbf{x}}(t)\delta t + \frac{\ddot{\mathbf{x}}(t)}{2}(\delta t)^2 + O((\delta t)^3) \tag{2.6a}$$

$$\dot{\mathbf{x}}(t+\delta t) = \dot{\mathbf{x}}(t) + \ddot{\mathbf{x}}(t)\delta t + \dots \tag{2.6b}$$

$$\ddot{\mathbf{x}}(t+\delta t) = \ddot{\mathbf{x}}(t) + \dddot{\mathbf{x}}(t)\delta t + \dots \tag{2.6c}$$

where $O(f(x))$ is notation for upper bound for absolute value of unknown or neglected component (error) of function.

This approach has several problems (cumulating errors, low numerical stability, etc.) which will not be described here in detail. More accurate methods of integration include the *Verlet algorithm* which in its basic form can be derived from the previous one (Equation 2.6a):

$$\mathbf{x}(t+\delta t) = \mathbf{x}(t) + \mathbf{v}(t)\delta t + \frac{\mathbf{a}(t)}{2}(\delta t)^2 + \frac{\mathbf{b}(t)}{6}(\delta t)^3 + O((\delta t)^4) \tag{2.7a}$$

$$\mathbf{x}(t-\delta t) = \mathbf{x}(t) - \mathbf{v}(t)\delta t + \frac{\mathbf{a}(t)}{2}(\delta t)^2 - \frac{\mathbf{b}(t)}{6}(\delta t)^3 + O((\delta t)^4) \tag{2.7b}$$

$$\mathbf{x}(t+\delta t) = 2\mathbf{x}(t) - \mathbf{x}(t-\delta t) + \mathbf{a}(t)(\delta t)^2 + O((\delta t)^4) \tag{2.7c}$$

In this form, the Verlet algorithm does not contain an explicit velocity, which simplifies calculations, but can be a problem in some MD simulations, when velocities are needed to calculate some system properties like temperature or kinetic energy. This problem is dealt with by calculating velocities using positions:

$$\mathbf{v}(t) = \frac{\mathbf{x}(t+\delta t) - \mathbf{x}(t-\delta t)}{2\delta t} + O((\delta t)^2) \tag{2.8}$$

(note the higher error, which is acceptable as it's not accumulated), or by employing more advanced versions of the algorithm (not discussed here).

In all numerical algorithms, the integration error is an increasing function of timestep. In addition to this, some processes put an upper limit on timestep — e.g. molecular vibrations frequency of 10^{12}–$10^{14} Hz$ require timestep of order of $10^{-15} s$ or $1 fs$ to properly observe their dynamics. On the other hand too short a timestep will cause longer calculation times and higher costs without significantly increasing precision.

2.3. Coarse grained simulation

Coarse grained methods are, in a way, the next step of approximation, found on the opposite side of *all-atom* molecular dynamics to that of quantum mechanics. Instead of dealing with single atoms, groups of atoms are treated together as *superatoms*.

In the simplest version, only hydrogen atoms are disregarded as separate particles and treated with heavier atoms, e.g. three atoms of CH_2 are treated as one heavy *pseudoatom* C_{2H} instead. This method is specifically called *united atom* simulation and was used extensively in earlier days of MD. Its importance as an approximation of all-atoms MD has diminished over time along side with improvements in computing technology.

For higher level coarse graining, groups of several heavier atoms are merged into superatoms like PO_4^{3-}, –COOH, $4\,H_2O$ etc. In popular

implementations of this method amino-acids are typically built of 2–5 superatoms each.

In all these method variants, the superatoms have to have their properties derived from their constituent atoms.

The biggest advantage of this method is the decreased size of model, which allows larger simulated times and larger systems for lower computational costs. Coarse grained simulations make it possible to capture higher level properties and processes like the formation of lipid bilayers. The price for this is obviously a lower resolution of the results.

2.4. Potential parametrisation

Because of the very large number of potential interactions, it is practically impossible to construct a force field with universal applicability. Therefore subsets of possible objects of simulations are chosen as a first step (e.g. inorganic crystals, proteins, nucleic acids) depending on the desired subject of research.

A quite common category of problems with existing force fields comes from the limited subset of parametrised interactions. Many biochemical structures contain atypical metal ions or inorganic molecules not considered notable enough to include in expensive parametrisation. These type of cases necessitate reverting to quantum mechanics (section 2.1) or other methods to compute the missing parameters before running any simulations.

Amber force field used throughout research for this thesis is described by the Equation 2.9. Bonding potential components are expressed in term of variables for bond lenght r, angle θ and dihedral angle ϕ with equilibrium values r_{eq}, θ_{eq} and γ respectively, and harmonic constants K_r, K_θ and V_n. Nonbonding interactions contain distances R_{ij} between unbound atoms, to compute *van der Waals* interaction (as Lennard-Jones potential 6–12) and *Coulomb* interactions (function of partial charges q_i, q_j),

with constant parameters A_{ij} and B_{ij}. Because calculating nonbonding interactions between all pairs of atoms is not practical for large models, atoms outside of the specified range are **cut-off** from computation.

$$V = \sum_{\text{bonds}} K_r(r - r_{eq})^2 \tag{2.9a}$$

$$+ \sum_{\text{angles}} K_\theta(\theta - \theta_{eq})^2 \tag{2.9b}$$

$$+ \sum_{\text{dihedrals}} \frac{V_n}{2}[1 + \cos(n\phi - \gamma)] \tag{2.9c}$$

$$+ \sum_{i<j} \left[\frac{A_{ij}}{R_{ij}^{12}} - \frac{B_{ij}}{R_{ij}^{6}} + \frac{q_i q_j}{\epsilon R_{ij}} \right] \tag{2.9d}$$

Each term of equation thus includes not only the actual state of the molecule in generalised coordinates, but also the state of reference (usually of minimum energy) and scaling constants. Equilibrium values, as well as other parameters, have to be derived by other means (e.g. quantum mechanical, experimental, thermodynamical) in a process called force field parametrisation, for each type of atom, which do not necessarily map one-to-one to chemical elements, except for the most simplified models. Typically, equilibrium values come from experiment and high-level *ab initio* calculations; the force constants are optimised to reproduce experimental and high-level *ab initio* vibrational frequencies of bonds and energy differences of conformations and rotational profiles[68]. Additional complication of this model comes from introducing different types of atoms depending on the chemical environment, e.g. aromatic vs aliphatic carbon. Amber forcefield employs 35 basic atom types, including five carbon, eight nitrogen, three oxygen and six hydrogen, as well as 22 special types[68].

2.5. Statistical methods

2.5.1. Trajectory interpretation

The immediate result of MD simulation is a set of trajectories containing positions and possibly velocities of all atoms over time. The first analysis may thus be presenting the trajectories in video or tabular format, but the usefulness of this approach is rather limited. It is more typical to employ mathematical methods of analysis. The simplest formal trajectory analyses are computations of statistical properties root mean square deviation (RMSD) and root mean square fluctuation (RMSF).

In RMSD (Equation 2.10a) the positions of all atoms or their significant subset (chosen molecules or types of atoms) are averaged for each time step and interpreted as a function of time. It is typically used as an indication of structure stability over time or the equilibration of a dynamic system.

In RMSF (Equation 2.10b) the same positions are averaged over the entire simulation time (or notable periods of time) for each atom (or group of atoms like molecule, residue) separately and interpreted as a function of atom (molecule, residue, etc). It may be used as a simple measure of the relative stability of the different parts of the molecule (e.g. protein domains), among others.

$$RMSD_t = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(x_i(t) - \tilde{x_i}\right)^2} \tag{2.10a}$$

$$RMSF_i = \sqrt{\frac{1}{T}\sum_{t=1}^{T}\left(x_i(t) - \tilde{x_i}\right)^2} \tag{2.10b}$$

Both functions are completely agnostic of the nature of the system described and have applications outside of molecular dynamics, as opposed to the specifically physicochemical functions described in section 2.6. While it is perfectly possible to modify the equations e.g. by taking

the weighted mean of atom masses or charges, this is rarely done, and this method is only used for crude initial analysis.

2.5.2. Correlation coefficient

Pearson's correlation coefficient (Equation 2.11) is calculated for two ordered sets of same cardinality (or equivalently a set of pairs) to test the *null hypothesis*. The sets are deemed to be uncorrelated if the coefficient is not significantly different from zero.

$$r(X,Y) = \frac{\sum_i (x_i - \bar{x})\,(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}} \tag{2.11}$$

2.6. Binding Energy

One of the most important thermodynamic properties of a system is its free energy. The thermodynamic free energy, or an energy which can be used to do non-mechanical work, is a state function, defined as a system's internal energy less the unusable entropic energy. Depending on experimental conditions, several different definitions of available energy are used. It may depend, among others, on temperature (T) and pressure (p), and indirectly on number (N) of molecules and volume (V) through gas equation. Biochemical reactions occurring in living organisms are usually interpreted in isothermal-isobaric (NPT) ensemble, with constant total number (N) of molecules, but also temperature (T) and pressure (p).

$$\text{Gas equation} \qquad pV = nRT \tag{2.12a}$$

$$\text{Internal energy} \quad U(S,V) \tag{2.12b}$$

$$\text{Enthalpy} \quad H(S,p) = U + pV \tag{2.12c}$$

$$\text{Helmholtz free energy} \quad A(T,V) = U - TS \tag{2.12d}$$

$$\text{Gibbs free energy} \quad G(T,p) = H - TS = U + pV - TS \tag{2.12e}$$

Free energy is a very useful concept for analysing reactions. The system's free energy is at a minimum when the system reaches equilibrium state. It is therefore convenient to use the difference in free energy between states as a criterion for spontaneous reactions. In this work the term "free energy" will henceforth imply *Gibbs* free energy (Equation 2.12e).

We define binding energy as follows:

$$\Delta G = G_{complex} - (G_{protein} + G_{ligand}) \tag{2.13}$$

Direct measurement of free energy is a non-trivial task and advanced methods are employed to achieve that. Advanced methods like Free energy perturbation (FEP) or Thermodynamic integration (TI) will not be discussed here in detail. Their accuracy comes at very high computational price. The focus will be put on approximate methods of determining free energy instead.

2.6.1. Continuum solvation methods

Two similar methods of molecular mechanics Poisson–Boltzmann surface accessibility (MM-PBSA) and molecular mechanics generalised Born surface accessibility (MM-GBSA) calculate binding free energies for macromolecules by combining molecular mechanics calculations and continuum solvation models[71]. Unlike more expensive exact methods, they can be used to calculate free energies from a single run of a receptor with ligand(s) in a complex, decreasing the cost of simulation at least threefold. Also, unlike more complex methods, they only require running a simulation of the immediate neighbourhood of the final (bound) state.

In practice, a larger number of simulations (replicas) is run simultaneously (the so called *ensemble* approach) instead of one very long simulation to achieve better coverage of the immediate neighbourhood in a states' phase space. Although MD itself is perfectly deterministic, both

thermostats and barostats used in simulations are usually stochastic, providing required variance between replicas.

The method involves running molecular dynamics simulations for bound complex in water solution with the total charge neutralised by counterions. Collected trajectories are then postprocessed by removing solvent together with the counterions, and calculating the energy components:

$$\langle G \rangle = \langle E_{MM} \rangle + \langle G_{sol} \rangle - TS_{MM} \tag{2.14a}$$

$$\langle G_{sol} \rangle = \left\langle G_{PB/GB} \right\rangle + \langle G_{SA} \rangle \tag{2.14b}$$

$$\langle E_{MM} \rangle = \langle E_{bonding} \rangle + \langle E_{vdW} \rangle + \langle E_{coulomb} \rangle \tag{2.14c}$$

where an electrostatic (polar) component $G_{PB/GB}$ of solvation energy is calculated using the Poisson-Boltzmann or Generalised Born equation respectively, and nonpolar G_{SA} is estimated by solvent accessible surface area. TS_{MM} is the solute entropy, which can be estimated by quasi harmonic or normal-mode analysis, but it is often ignored due to its high computational cost, compared to relatively low contribution to total energy[72].

By substituting G from Equation 2.14a in Equation 2.13 bonding components cancel out, the entropic component being very small in comparison is therefore ignored, thus giving the final equation:

$$\Delta G = \Delta \langle G_{sol} \rangle + \Delta \langle E_{vdW} \rangle + \Delta \langle E_{coulomb} \rangle \tag{2.15}$$

MM-PBSA

The MM-PBSA method of calculating binding energy is an approximate method that has been proven successful in free energy calculations in large number of experiments. With relatively good coverage of a system states' phase space, reasonably accurate results can be obtained.[72].

MM-GBSA

The MM-GBSA method is very similar to the previous one. The only difference is in calculating polar component of solvation energy by Generalized Born equation (G_{GB}). The comparisons between both methods are numerous and inconclusive. The MMGBSA is generally accepted as underperforming in calculating absolute binding free energies, compared to MM-PBSA[71]. It is nevertheless used extensively because of its lower computational cost and decomposability. Binding energy per residue decomposition is often essential in identifying structural determinants of the binding affinity[73].

2.7. Thermodynamic view of binding affinity

As elaborated earlier, RLT is the Integrase inhibitor binding the active site competitively to DNA which would otherwise be integrated during the strand transfer phase of HIV lifecycle.

This work is not in any measure extensive research on thermodynamics, but some basic concepts will have to be introduced for easier reception of the motivations and results of this project.

Most biological processes, including HIV strand transfer catalysed by Integrase, are on a basic level chemical reactions between large molecules. The enzyme A (Integrase) binds its ligand B (DNA) non-covalently in a reversible reaction before performing the actual enzymatic reaction.

$$A + B \underset{k_d}{\overset{k_a}{\rightleftharpoons}} AB \xrightarrow{k_e} A + B^* \tag{2.16a}$$

$$K_a = \frac{1}{K_d} = \frac{k_a}{k_d} = \frac{[AB]_{eq}}{[A]_{eq}[B]_{eq}} \tag{2.16b}$$

where k_a and k_d are rate (kinetic) constants of association and dissociation respectively, k_e is rate constant of essentially irreversible enzymatic reaction, K_a and K_d are equilibrium constants defined by equilibrium concentration $[AB]_{eq}$, $[A]_{eq}$ and $[B]_{eq}$. This very simplified picture ig-

nores the complexity of catalysed enzymatic reaction step, to focus on the first step only, which is the target of the inhibitor. The competitive inhibitor is an alternative ligand which, when binding to the enzyme blocks (inhibits) the enzymatic reaction. The inhibitor may exist in a natural system as part of physiological pathway or in the case of most drugs be engineered to maximise the binding affinity and subsequently the efficacy. The exact mechanism of inhibition may vary tremendously, and in the case of HIV Integrase inhibitors is not fully explained, but the effects on system dynamics may be approximated by equations analogous to Equation 2.16a and Equation 2.16b:

$$A + B + I \underset{k_j}{\overset{k_i}{\rightleftharpoons}} AI + B \tag{2.17a}$$

$$K_i = \frac{k_i}{k_j} = \frac{[AI]_{eq}[B]_{eq}}{[A]_{eq}[I]_{eq}[B]_{eq}} \tag{2.17b}$$

where K_i is inhibition equilibrium constant measured at equilibrium concentration of enzyme-inhibitor complex ($[AI]_{eq}$), free enzyme ($[A]_{eq}$) and free inhibitor ($[I]_{eq}$). The concentration of a ligand ($[B]_{eq}$) which does not participate in the reaction cancels out.

Combining Equation 2.16b and Equation 2.17b by solving and substituting concentration of free enzyme — which is in equilibrium with both complexes — gives a ratio of a protein bound to ligand (involved in a first step of enzymatic reaction) to inhibited protein as a function of both thermodynamic constants and concentrations of competing ligands:

$$\frac{[AB]_{eq}}{[AI]_{eq}} = \frac{K_a[B]_{eq}}{K_i[I]_{eq}} \tag{2.18}$$

Equilibrium constant is related to Gibbs binding free energy introduced in section 2.6 through reaction isotherm equation:

$$\Delta G = -RT \ln(K_a) \tag{2.19}$$

Finally, the thermodynamic constant from Equation 2.19 can be substituted into Equation 2.18 giving the ratio of inhibited protein in terms of free energy levels of involved states.

$$\frac{[AB]_{eq}}{[AI]_{eq}} = \frac{[B]_{eq}}{[I]_{eq}} \exp\left(-\frac{\Delta G_B - \Delta G_I}{RT}\right) \tag{2.20}$$

The Equation 2.20 shows the relation between binding energies ΔG_I calculated throughout this project and the inhibition efficiency of a drug. The subscript I is conveniently omitted throughout the rest of this thesis but it is clear from the context when calculated values of ΔG refer to ΔG_I, where complex of In with RLT is analysed. The complex of Integrase with inhibitor showing lower (more negative) ΔG_I means less In bound to DNA, or more effective inhibition. Conversely, higher (less negative) ΔG_I means less effective inhibition or more resistant viral enzyme.

3. Homology Modelling of HIV integrase

In this chapter I will describe the attempts to obtain a structure of HIV IN by the methods of Homology Modelling. As one of the desired structure purposes is further MD experiments, a high resolution all-atom structure is required, hopefully of a complete protein complex with the drug and the DNA attached, to simulate the natural process of DNA binding inhibition by the drug.

3.1. Introduction

Homology modelling is one of the methods used for detailed prediction of protein structures which cannot be obtained by experimental methods directly. Is is based on the fact that all proteins are coded by genes which are often closely related. The closer the evolutionary relationship between the protein coding genes – the more similar the protein structure and function likely are. By using several homologous proteins as structure templates one may be able to predict the unknown structure with reasonable precision.

3.1.1. Proteins structure

Protein molecules consist of one or more polypeptide chain built of amino acids residues by peptide bond (Figure 3.1), and often but not necessarily other atoms and molecules (ligands) including metal ions, water molecules, acid residues (phosphates, sulphates, etc.) and others.

Most enzymes' functionality involves binding and releasing the ligands in the catalysed processes which adds another level of complexity.

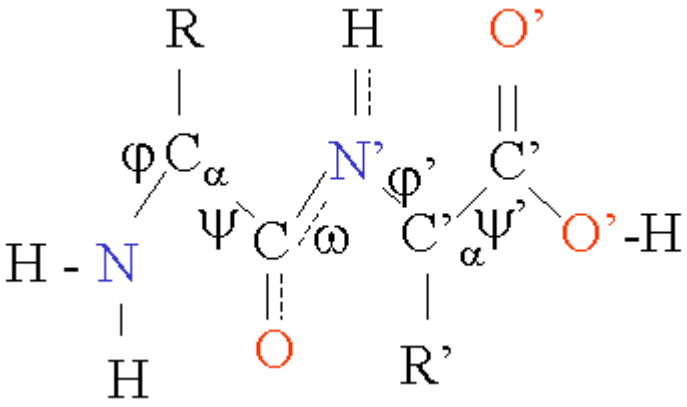

Figure 3.1. The Peptide bond is a type of amide bond formed between carboxylic group of first amino acid and amine group of second one. Amino acid sidechains are replaced with symbols R and R' respectively for simplicity.

There are 20 standard biological amino acids and a number of specific amino acids modified in post-translation process. All amino acids have the same basic formula $NH_2{-}C_\alpha(R){-}COOH$ where R denotes a residue side chain, with the exception of proline which side chain form a heterocycle including both C_α and $-NH-$ as well as three $-CH_2-$ groups. While all amino acids are different from each other in some respect or another they can be classified in several subcategories by their various physicochemical properties.

Primary structure is the sequence of amino acid residues in a peptide chain. It is usually described by listing the one letter codes for amino acid residues starting from N-terminus (amino acid with free $-NH_2/-NH_3^+$ group) up to C-terminus (amino acid with free $-COOH/-COO^-$ group) i.e. the order it is assembled *in vivo*. The primary structure can be easily discovered by sequencing the protein gene or the protein molecule itself. Both these methods have been used routinely for some time now.

Changing one amino acid for another in a protein sequence may have effects ranging from complete loss of protein function to completely neg-

ligible, depending among other factors on preservation of amino acid properties.

Secondary structure is the first level of three-dimensional organisation of protein molecule. The most characteristic secondary structures are α-helix and β-sheet suggested for the first time by Astbury[74] and described in detail by Pauling[75]. In the α-helix backbone CO groups form hydrogen bonds with groups NH four residues further. In the β-sheet the hydrogen bonds are formed between backbone groups in parallel or anti-parallel chains. There are other known secondary structures and most proteins also possess regions of no particular secondary structure.

Tertiary structure is the name given for high order three-dimensional molecule organisation. Helices and sheets, as well as coils and unordered chains fold into compact molecule (Figure 3.2). By bringing distant

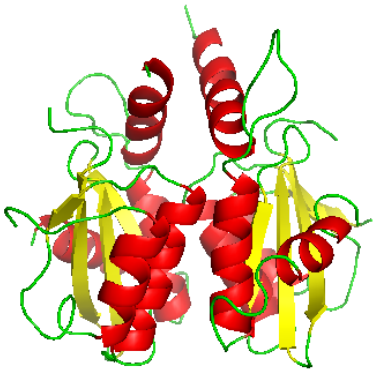

Figure 3.2. Protein structure on example of SIV IN. Typically protein sidechains are omitted and backbone chain is shown schematically as solid spirals (α-helix shown in red), arrows (β-sheet in yellow) or strings (sections lacking secondary structure, in green). The image shows higher level of structure by depicting two protein molecules of two domains each.

parts of the molecule together an active site may be formed and give the molecule ability to perform its function. The process of folding is the "Holy Grail" of computational structural biochemistry. It has been shown that any random process of looking for optimal conformation as

well as systematic search of whole conformations space would take much longer than the age of the universe, even for simple proteins[76]. Thus only some heuristic methods of predicting protein folding are known and they are far from successful.

Ternary or Quaternary structure names are given to higher orders of arrangement; depending on context it may mean a spatial relationship between domains of one molecule or of molecules constituting an oligomer.

3.1.2. Structure determination

The most popular method of determining biomolecules structures deposited in Protein Data Bank are —in decreasing order— X-ray crystallography, Nuclear Magnetic Resonance and Electron Microscopy (Table 3.1).

Method	Proteins	Nucleic Acids	Prot+NA	Other	Total
X-ray	58 735	1 268	2 844	18	62 865
NMR	7 709	943	169	7	8 828
EM	250	22	92	0	364
hybrid	28	3	1	1	33
other	132	4	5	13	154
total	66 854	2 240	3 111	39	72 244

Table 3.1. PDB structures breakdown by methods and type of molecule —as of 5 Apr 2011— according to RCSB[77].

Each of the methods has its advantages and disadvantages, but many protein structures have eluded determination by any of them for years. In these cases theoretical methods are employed to *predict* instead of *determine* the structure.

3.1.3. Phylogenetics

One of first steps required for successful homology modelling is establishing degree of homology between available structures. The most

closely related homologue structures will subsequently be used in the modelling.

Protein homology may be result of either a speciation (orthology) or a gene duplication within organism (paralogy), but in case of viruses the latter in very unlikely thus the term homology will subsequently imply orthology or relationship between proteins and genes from related species. It is generally agreed that the closer the homologue the better it serves as modelling template. To that end the relationships between known structures is analysed using phylogenetic analysis. Unrelated proteins often share some structural and/or functional properties e.g. due to evolutionary convergence but they are usually useless as modelling templates because their similarities are superficial. Different strains of HIV, e.g. had been reported to convergently evolve nuclear receptor-responsive element (NRRE) in the LTR region[78]. On the other hand closely homologous protein with unrelated function may still share fundamental structural priorities with the protein of interest and be useful as a template.

During the course of evolution protein sequences change through point mutations of genetic code. Due to redundant nature of the code many mutations don't affect the expressions at all, mutation in 3rd nucleotide of a codon has high chance of preserving encoded amino acid. Of those mutations that do replace one amino acid with another not all have measurable effect on protein properties thanks to the fact that similar codons often encode amino acids with similar properties.

Amino acids share many properties between them and thus replacing one with very similar may in fact have negligible effects. Similarly, change in relatively unimportant region of a protein, far from active centre may have no effect on its function at all unless it causes structural change or is responsible for essential interaction with environment. Such neutral point mutations, while meaningless for the organism's biochemistry are useful

in research of genealogical relationships between species and individuals, because they are not subject to evolutionary selection.

On the other hand mutation may have dramatic effect if changed amino acid had performed catalytic or important structural role. Typical examples include Sickle-cell anæmia caused by mutation E6V in β-globin gene (meaning sixth position in a sequence, normally occupied by Glutamic Acid is replaced with hyrophobic Valine, changing overall protein shape and affecting its function as a result) or Hæmochromatosis (H63D or C282Y in HFE gene[80][81]).

Some mutations can be potentially beneficial to their carriers e.g. by improving enzyme efficiency, providing better protection from diseases, new functionality. In case of protein being drug targets, mutation may reduce drug efficiency by changing binding site structure, change beneficial for targeted organism but detrimental for therapy prognosis for a patient.

The immediate goal behind phylogenetic analysis of a group of related sequences is to infer a degree of relationship between them, then usually presented in form of binary tree. The phylogenetic tree represents the evolutionary history of a group where bifurcations represent speciacion or gene duplication events. Most methods also allow estimation of relative degree of divergence which —with sufficient temporal data from other methods like history or palæontology— may be used to estimate dates of all bifurcation events.

Successful analysis requires fulfilling several conditions:

— Evolutionary events are strictly bifurcating with no horizontal transfers

— Each position in sequence evolved homogeneously and independently

— The sequences are correct and from specific sources

— The sequences are actually closely related

— The sampling data is sufficient and adequate

Two first conditions are completely independent from the researcher and are rather assumptions about the subject of research. The conditions on input data, while often hard to verify beforehand can be improved through additional research and more hands-on iterative analysis. Detailed initial analysis may in particular show insufficiency of data, lack of identifiable relationship between some sequences or even hint at convergent evolution. In these cases revision of input set is required.

3.2. Methods and tools

3.2.1. Sequence alignment and Clustering

A systematic analysis of relationship was attempted for HIV-1 IN and several homologous proteins for which structures had already been solved — HIV-2 (3F9K), SIV (1C6V), MVV (3HPG), RSV (1C0M), ASV (1CXQ) and PFV (3L2Q) — as found in Protein Data Bank (PDB), with their respective PDB codes in parentheses. The names and structures behind these acronyms will be revealed in section 3.3. The genealogical relationship was reconstructed by the progressive sequence alignment analysis with the ClustalW tool[82]. Progressive, or hierarchical alignment method combines clustering and multiple sequence alignment into an iterative process during which a *guiding tree* is constructed using default Neighbour Joining method[83], producing an *annotated alignment*. The graphical representation of phylogenetic tree was generated with the program Archæopteryx[84] from the annotated alignment.

3.2.2. Modelling

Having obtained a phylogenetic tree with sequence alignment the next step is to use this information to predict the structure of the protein in question. The modelling relies on the assumptions that protein structure is determined by a sequence, hence similar sequences yield similar structures. Known protein structures defined by aligned sequences are

therefore aligned in space in a way that minimises distances between homologous amino acid residues. At this stage some proteins may turn out not to preserve the structure of their homologues and may need to be discarded. Aligned structures are then used as a three-dimensional template for the sought structure, while phylogenetic distances may be used as weights in dubious sections.

Swiss-PdbViewer —also known as DeepView— is one of the most popular tools for structural bioinformatics[85]. It provides an integrated sequence to structure analysis and modelling platform[86]. DeepView was used for structural alignment of overlapping protein fragments, introducing mutations, filling short gaps ("loop building") and other model building tasks, as well as parallel orientation of structures for visualisation purposes and short energy minimisation using built-in Gromos96 force-field.

Modeller software[87] is designed for minimal researcher intervention which puts it in the category of "Black Box" tools. In the simplest case providing template structures and sequence alignment is enough to generate a model. However, further modifications and optimisations are usually required to obtain useful structures. Because of the "Black Box" nature of the process it is difficult to evaluate the final model without comparing it to the real structure, which is usually unknown at this stage. It may, however, be combined with other data, e.g. as a template for Molecular Replacement method for x-ray crystallographic experiment.

3.3. Input data

No complete structure of HIV IN, with or without substrates, is known as of today. Unfortunately, all attempts to crystallise HIV-1 IN turned out to be unsuccessful. Similarly structure determination by the method of NMR is thwarted by low solubility of IN. However, several

homologous proteins, as well as partial structures have been solved. All protein molecules henceforth are visualised with PyMol molecular viewer, render tool, and 3D molecular editor[88]. Where appropriate, PDB structure of visualised molecule is indicated by its four-character code.

3.3.1. HIV-1 Integrase Homologues

Some lentiviruses had already been researched even before HIV discovery. Several structures homologous to HIV-1 IN have been solved and deposited in PDB giving insight into the common features of the lentiviral integrases family. As all those structures have a potential use in homology modelling the review of available data is presented below.

ASV integrase

Figure 3.3. ASV Integrase, PDB:1CXQ

The crystallisation of Avian Sarcoma Virus (ASV) integrase CCD (residues 52–198) monomer under a range of conditions had shown the high flexibility of active-site loop 144–154 (homologous with HIV-1 loop 138–148), suggesting functional conformational changes depending strongly on pH[89]. The structure of ASV IN is shown in Figure 3.3.

RSV integrase

Following the research showing that two out of three domains (CCD and CTD) of RSV together are able to perform terminal cleavage and strand transfer[90] the crystallographic structures of RSV IN (res 49–286)

Figure 3.4. RSV Integrase, PDB:1C0M

and its mutants were solved[91]. Notably, only slightly shorter RSV IN (58–286) had shown no correct activity. Despite not being directly involved in active site ions binding, the missing segment (49–57) is close enough to it to affect the process, and it lies in the probable DNA binding surface. The structure is shown in Figure 3.4.

SIV integrase

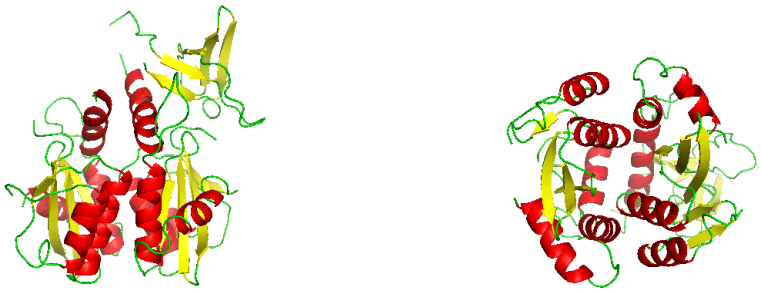

Figure 3.5. SIV Integrase, PDB:1C6V

The structure of two-domain SIV IN even closer HIV IN homologue confirmed the CCD and CTD structure, as well as the domains relative position[92]. At the time only single domains of HIV IN were solved, thus the structure of such a close homologue was an important indication of a DNA binding site along both domains. The structure shows the asymmetric dimer of dimers as indicated for active form by earlier works[93]. See the structure of a dimer in Figure 3.5.

MVV integrase

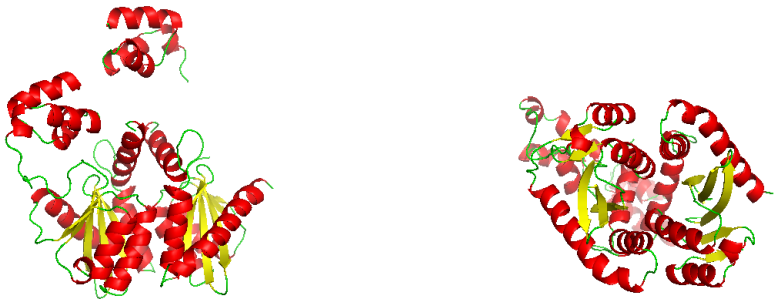

Figure 3.6. MVV Integrase, PDB:3HPG

Two-domain Maedi-Visna Virus (MVV) (4–212) structure with its cofactor Lens epithelium-derived growth factor (LEDGF) shows possible tetrameric configuration of active complex[94]. The structure of a single dimer is shown for comparison in Figure 3.6.

HIV-2 integrase

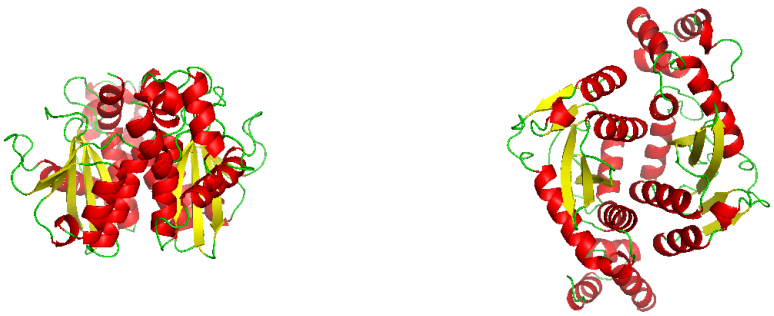

Figure 3.7. HIV-2 Integrase, PDB:3F9K

The structure of NTD and CCD HIV-2 IN domains (1–207) with LEDGF sheds further light on the question of charge based interactions between the constituents of active complex[95]. The structural and functional similarity between HIV-1 and its closest homologue HIV-2 have been confirmed by engineering HIV-1 mutant with increased affinity to LEDGF Integrase Binding Domain (IBD) using HIV-2 structure. The structure of HIV-2 IN dimer is shown in Figure 3.7.

PFV integrase

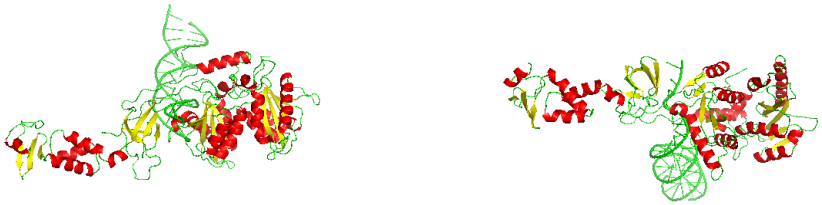

Figure 3.8. PFV Integrase, PDB:3L2Q

The first structure of integrase complete **intasome** shows complete PFV IN molecule interacting with 19-bp (base pairs) DNA[96]. The DNA seems to be coupled tightly with IN active site, with the 5'end 3-bp (ATT) heavily distorted from its native conformation and separated from its complementary 3'end A – remaining two bp being already cleaved out.

Both N-term domain stabilising Zn–HHCC and core Mg–DDE motifs are present, although due to poor resolution Glu221 does not seem to take part in chelating Mg. Instead, the structure with Mn shows more unambiguous complex of **two** Mn ions with D,D(35)E. The Mn/Mg ions in the complex are homologous to the Cd/Zn in ASV, illustrating the consistent organisation of active site regardless of type of ion.

The structures of IN intasome with drugs MK0518 (RLT and GS9137 (ELV) were solved in the same set of experiments. The drugs seem to fit well within the active site and not only interact with metal–DDE complex but also bind and displace the DNA 3'end A, possibly explaining the mechanism of inhibition.

The significance of experimental data regarding IN intasome is countered by the fact that PFV is the most distant homologue of HIV-1, out of several reviewed in this work. Although the active site is very well conserved among all of them, the PFV IN overall high-level conformation, highly dependent on the domains internal structure is notably different. The

exact mechanism of DNA binding by HIV-1 is thus still a matter for future research. The usability of PFV IN intasome structure in simulating HIV-1 IN intasome model dynamics still needs to be determined.

Summary

The catalytic core domain is clearly very well conserved among all homologues. The structure of the active site seems not to be significatly affected by mutations introduced to increase solubility and/or crystallisation. The remaining domains show more variation in both sequence and structure (when it is known). Nevertheless, all data gathered through research on HIV-1 IN homologues contributes to the general knowledge of Integrases structure and function.

3.3.2. HIV-1 Integrase partial structures

Although complete HIV-1 structure has not been crystallised yet, several partial structures have been solved instead. Partial structures, especially when combined with complete homologues are valuable as an input for homology modelling.

Core domain

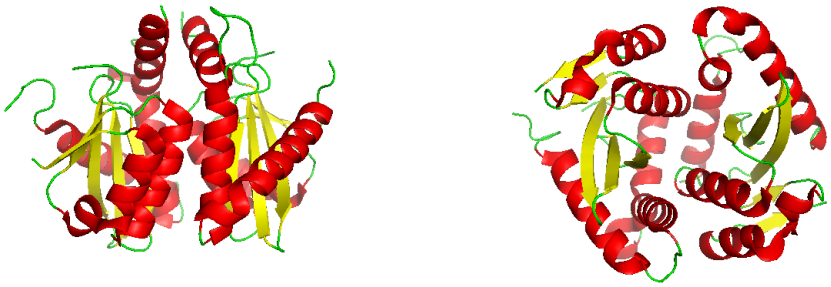

Figure 3.9. HIV Integrase core domain, PDB:1QS4

The first structure of HIV-1 IN core domain (55–209) was solved in 1994 (pdb entry 1ITG)[97], followed by solution of active site binding Mg ions in ASV-IN (1ASU, 1ASV, 1ASW)[98] and HIV1-IN (1BH2, 1BI4)[99]

which confirmed the high level of similarity between family members, at least in respect to the active site, specifically D,D(35)E motif. Finally the structure of core domain with bound inhibitor was solved in 1999 (1QS4)[100].

Core and C-term domains

Figure 3.10. HIV Integrase CCD and CTD, PDB:1EX4

The first multidomain structure of HIV-1 IN, solved in 1999, was of a five point mutant core and C-term domains (56–270) [101]. A short segment of core domain is missing (142–144) from the pdb structure. The structure shows a Y-shape dimer. Core domains form surface interaction between molecules while C-term domains remain 55Å apart. The dimerisation through core domain shows similarity to both RSV[91] and SIV[92] integrases which suggests it will also be found in full-length IN[101].

The relative arrangements of domains, however, differs significantly between homologues, suggesting flexibility. The link between core and C-term domains of HIV IN shows some flexibility around T210 and introduces asymmetry to the dimer by 90° rotation of one of the C-term domains[101].

A contiguous strip of positive charge along the outside surface of the dimer suggests the binding path for DNA[101].

Figure 3.11. HIV Integrase NTD and CCD, PDB:1K6Y

N-term and core domains

A structure of N-term and core domains confirms the dimeric structure again[102]. The direct interaction between respective CCDs is found again, except the slight difference of one loop (188–194) on the interdomain interface. The link between domains (47–55) is missing, as well as part of the core domain (140–148).

The N-term domains are stabilised by one Zn ion each, chelated by His12, His16, Cys40, Cys43 (HHCC motif). The relative spatial orientation of N-term domains in the two domains structure differs significantly from that of isolated domains due to the arrangement of core domain and interdomain link[102].

Once again asymmetry is introduced by interdomain link causing one of N-domains to be moved by 15° in relation to core domain[102].

The phosphate anion identified 7Å from away the active site is suggesting the binding site of DNA backbone[102].

The crystal structure shows higher order dimer-of-dimers which would confirm earlier suggestions that IN acts as a tetramer *in vivo*. Although the tetramer of HIV IN NTD–CCD protein was not observed in solution, only in crystal[102], the CCD–CTD protein tetramerises in solution as well[103] indicating a role of C-term domain in tetramerisation. It has also been pointed out that superimposing partial complementary structures of HIV-IN tetramer does not introduce any significant steric clashes

and shows some similarity to Tn5 transposase[102] leaving the possibility of DNA binding similar to that of Tn5[104]. The shape of the groove would require bending of DNA for integration, which is consistent with earlier findings[105][106][107].

3.3.3. Theoretical models

Several documented attempts have been made to determine HIV-1 IN by the means of theoretical methods. Although the generation of a feasible protein model by superimposing known partial structure is pretty straightforward[102], with only a few residues left to modelling, the actual problem is the binding site of DNA.

The first attempts involved manual docking of DNA into the protein[108][109] or protein–drug complex[110].

Anther approach involves automatic docking of DNA using a modified version of ESCHER program[111], normally used for protein–protein interactions.

3.4. Results

In this chapter will be described attempts to generate a plausible model of HIV-1 IN based on existing partial and homologous structures. With an actual structure unavailable or incomplete, homology modelling may provide a viable alternative. Unfortunately, without high quality structures of close homologues the quality of model may turn out not to be satisfactory for the purpose of MD simulations. However, even if not sufficient for detailed quantitative analyses, it may still be proven useful for qualitative insight into the working of the protein.

3.4.1. Sequence alignment

The genealogical relationship was reconstructed by multiple sequence alignment analysis with the ClustalW[82] command line tool. The alignment of the most conserved region is shown in Table 3.2.

Species	pdb	res-	sequence		-res
HIV-1	—	55	.DCSPGIWQLDCT..HLEGK	VILVAVHVASGYIEAEVIPA	91
HIV-2	3F9K	55	.NAELGTWQMDCT..HLEGK	IIIVAVHVASGFIEAEVIPQ	91
SIV	1C6V	55	.NSDLGTWQMDCT..HLEGK	IVIVAVHVASGFIEAEVIPQ	91
MVV	3HPG	47	.KRGIDHWQVDYT..HYEDK	IILVWVETNSGLIYAERVKG	93
RSV	1C0M	55	.LGPLQIWQTDFT..LEPRM	APRSWLAVTVDTASSAIVVT	91
ASV	1CXQ	55	.LGPLQIWQTDFT..LEPRM	APRSWLAVTVDTASSAIVVT	91
PFV	3L2Q	118	PQKPFDKFFIDYIGPLPPSQ	GYLYVLVVVDGMTGFTWLYP	157
alignment			.......+..*.........	.....+.:..:......+..	
HIV-1	—	92	ETGQET.....AYFLLKLAG	RWPVKTIHTDNGSNFTGATV	126
HIV-2	3F9K	92	ESGRQT.....ALFLLKLAS	RWPITHLHTDNGANFTSQEV	126
SIV	1C6V	92	ETGRQT.....ALFLLKLAG	RWPITHLHTDNGANFASQEV	126
MVV	3HPG	94	ETGQEFR.....VQTMKWYA	MFAPKSLQSDNGPAFVAEST	128
RSV	1C0M	92	QHGRVTSVAVQHHWATAIAV	LGRPKAIKTDNGSCFTSKST	131
ASV	1CXQ	92	QHGRVTSVAAQHHWATAIAV	LGRPKAIKTDNGSCFTSKST	131
PFV	3L2Q	158	TKAPSTS..ATVKSLNVLTS	IAIPKVIHSDQGAAFTSSTF	195
alignment			..:..................	....:.+++*+*:.*:::...	
HIV-1	—	127	RAACWWAGIKQEFGIPYNPQ	SQGVVESMNKELKKIIGQVR	166
HIV-2	3F9K	127	KMVAWWIGIEQSFGVPYNPQ	SQGVVEAMNHHLKNQISRIR	166
SIV	1C6V	127	KMVAWWAGIEHTFGVPYNPQ	SQGVVEAMNHHLKNQIDRIR	166
MVV	3HPG	129	QLLMKYLGIEHTTGIPWNPQ	SQALVERTHQTLKNTLEKLI	168
RSV	1C0M	132	REWLARWGIAHTTGIPGNSQ	GQAMVERANRLLKDRIRVLA	171
ASV	1CXQ	132	REWLARWGIAHTTGIPGNSQ	GQAMVERANRLLKDKIRVLA	171
PFV	3L2Q	196	AEWAKERGIHLEFSTPYHPQ	SSGKVERKNSDIKRLLTKLL	235
alignment			.......**....:.*.+:*	:::.**..+..+*..+..+.	

Table 3.2. Multiple sequence alignment of most conserved region of lentiviral integrases. Active centre D,D(35)E motif residues are shown in red (see subsection 1.4.3). Standard ClustalW scheme for results display - an asterisk (*) denotes residue conserved among all compared sequences, plus (+) and colon (:) indicate decreasing conservation, and dot (.) the remaining residues with insufficient similarity.

3.4.2. Phylogenetics

The phylogenetic tree was built with the program Archæopteryx[84] on the previously generated alignment. The approximate genealogical relationship between proteins inferred from their sequences is shown in the Figure 3.12.

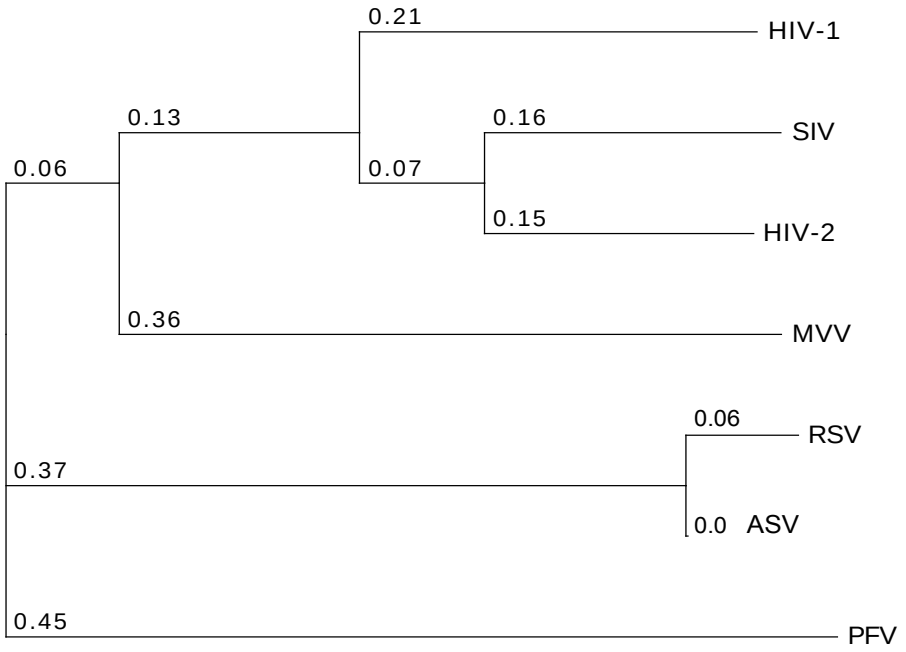

Figure 3.12. Phylogenetic tree of lentiviral integrases. Branch lengths proportional to amount of inferred evolutionary change between nodes and their ancestors. Branch labels describe ratio of change in relation to whole integrase sequence.

3.4.3. Modeller

The Modeller software tool[87] was used for model generation, using a previously generated sequence alignment with different sets of templates chosen from HIV-1 structures (1K6Y, 1EX4, 1WJA) and complete homologous PFV intasome (3L2Q). While closer homologues exist, their structures —being incomplete— do not bring much additional information required for model building. Tertiary structure of the molecule is quite consistent between all generated models, nevertheless the variance between structures are too big to ignore, decreasing confidence in their value as foundation for further research.

3.4.4. Initial model

The initial mock models of HIV Integrase were generated using three partial HIV-1 structures (1K6Y-A, 1EX4-A, 1WJA-A) and complete ho-

mologous PFV intasome (3L2Q-A). The model did not show any unusual features, thus encouraging taking a more advanced approach using the same technique.

3.4.5. Monomeric apoprotein molecule

The template for IN model was manually generated by superimposing 1K6Y-B and 1EX4-A core domains over 1QS4-B and then merging segments from all chains to obtain an almost complete template. There were still gaps in positions 47–55 (between NTD and CCD), 142 (in CCD) and 271–288 (the final segment of CTD).

HIV-1 IN sequence were used to replace mutated residues from the template with wild type residues. Some sidechains exhibited high energy due to interactions with reintroduced sidechains. They were manually reorientated before further optimisation. Missing residues were reintroduced to the structure by DeepView Loop Builder and the complete model energy was minimised. The resulting model of apomonomer is shown in Figure 3.13.

3.4.6. Modelling the dimer

Initially the dimer was modelled by superimposing the same generated monomer over both respective structures from 1QS4. There were areas of high energy interactions between molecules which were eliminated by simple minimisation. Figure 3.14 shows the dimer constructed from the identical monomers.

In the next approach the asymmetric dimer was built starting from 1QS4 using respective structures from 1K6Y and 1EX4. The difference in energy between symmetric and asymmetric structures were not significant at this stage, but it has to be noted that the substrate (DNA) is probably the stabilising factor of asymmetric structure. The resulting model is shown in Figure 3.15.

Figure 3.13. Single apoprotein model generated with DeepView by superimposing matching parts of 1EX4-A and 1K6Y-B core domains over 1QS4-B.

Figure 3.14. Symmetric dimer model generated with DeepView by superimposing two copies of previous model over 1QS4 dimeric structure.

Figure 3.15. Asymmetric dimer model generated with DeepView by superimposing matching parts of 1EX4 and 1K6Y core domains dimers over respective 1QS4 structures.

3.4.7. Challenges

While obtained dimeric apoprotein model is in agreement with other suggested structures, binding DNA and metallic ions lacks consistent support in literature ([108], [109], [110], [111]). Without a high confidence active centre model no quantitative description of drug binding is possible.

3.4.8. Conclusions

It has been decided to focus further research on well described crystallographic structures of homologous proteins instead of homology models of inconclusive reliability. Although the modelling of protein seems to provide reliable results the problem of docking DNA remains. The most difficult task, of modelling the DNA interaction with the IN active site cannot be dealt automatically by any software. With only one intasome structure available there is not enough data for generalisation. However, as the active site proves to be very well conserved among all homologues it is suggested that the drug binding mechanism may be properly described with satisfying accuracy basing on homologous structures.

4. PFV Integrase intasome binding energy in MD

As noted earlier, no complete crystal structure of HIV Integrase is currently available. However, several structures of Prototype Foamy Virus (PFV) intasome are solved with different ligands.

This chapter will describe the application of Molecular Dynamics (MD) to the original complete PFV Integrase (IN) intasome monomer structure and the subsequent computation of binding energies using the molecular mechanics Poisson–Boltzmann surface accessibility (MM-PBSA) and molecular mechanics generalised Born surface accessibility (MM-GBSA) methods to verify its usefulness as input data for calculations of drug binding energies in an attempt to explain strand transfer inhibition by Raltegravir (RLT). PFV Integrase intasome structure[96] (PDB:3L2T, as shown in Figure 4.1) with RLT in its active centre was chosen for all simulations as the most complete system available.

The structure contains a fragment of double-stranded DNA helix which *in vivo* would be the end of a longer HIV DNA to be integrated into the host genome. To emulate the restrained movement of such a configuration parts of DNA in this structure have had their movement restricted by applying harmonic constraints in some of the simulations.

In the input data, the protein molecule is bound with a DNA double helix, two magnesium ions and one RLT molecule in active centre (shown in Figure 4.2). One zinc ion is permanently bound by NTD as well, contibuting to its structure. The active site contains one copy of RLT, Integrase strand transfer inhibitor (INSTI) in a position indicating the

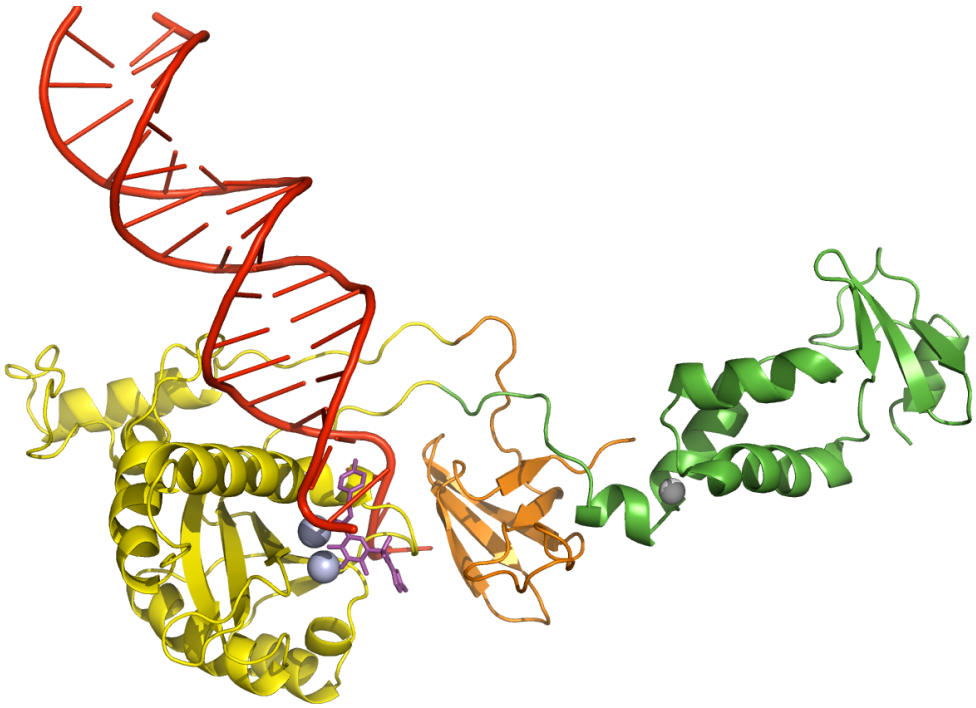

Figure 4.1. Complete structure of monomeric unit of PFV intasome with inhibitor (PDB:3L2T). Parts marked by colours — NTD (green), CCD (yellow), CTD (orange), DNA (red), RLT (magenta), Mg (light blue), Zn (white, in NTD shade).

competitive binding, displacing CA_{OH}-3' dinucleotide. The structure of RLT alone is shown in Figure 4.3.

4.1. Methods

The Amber force field, described in chapter 2, was chosen for all simulations, as it is known for relatively good support for nucleic acid molecules, without compromising its overall performance for proteins.

4.1.1. Input data and Simulation parameters

Raltegravir

RLT molecule charge was confirmed to be neutral by protonation state analysis using *Protonation* bundle of ChemAxon's *Physico-chemical property predictors* Calculator Plugin[112]. Gaussian 03 was used to per-

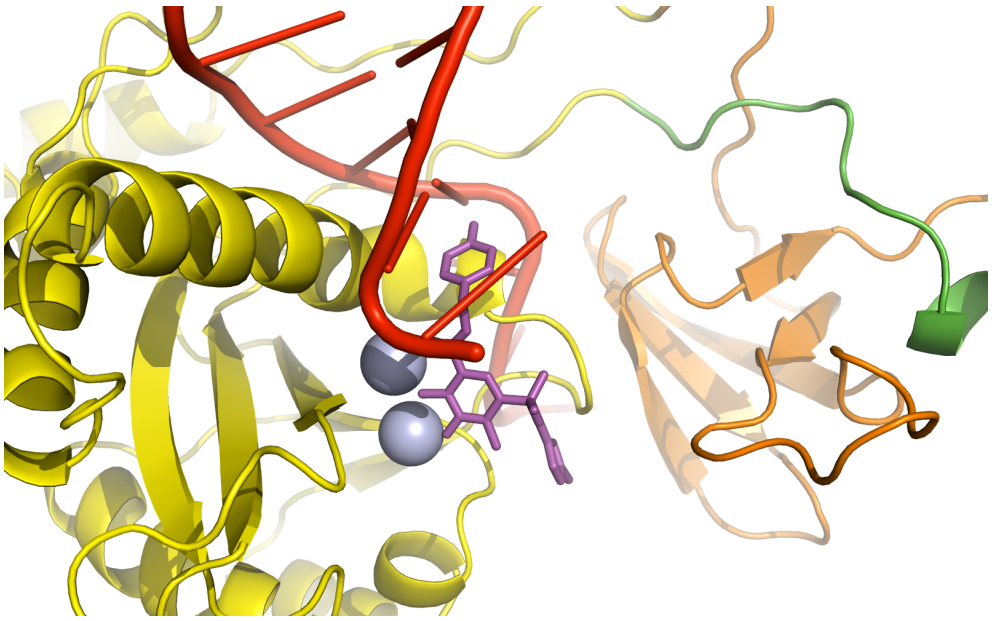

Figure 4.2. Active centre structure of PFV intasome with inhibitor. Parts marked by colours — NTD (green), CCD (yellow), CTD (orange), DNA (red), RLT (magenta), Mg (light blue). RLT is shown competitively binding IN active centre, inhibiting DNA binding.

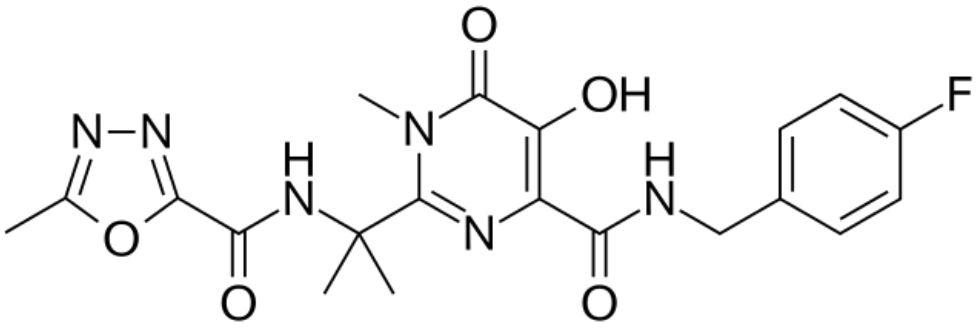

Figure 4.3. Structure of Raltegravir.

form geometric optimisation of the drug molecule using the Hartree–Fock method. Partial atomic charges were generated with the Restrained Electrostatic Potential (RESP) procedure. Antechamber toolkit was used to complete forcefield parametrisation of the drug in General Amber Forcefield (GAFF)[68].

Forcefield

Amber standard ff03.r1 forcefield was used to parametrise the Integrase receptor consisting of a protein molecule and double DNA helix[113].

The Zn^{2+} ion chelated by the tetrahedral HHCC motif (subsection 1.4.3) within the NTD was parametrised as Tetrahedron-Shaped Zinc Divalent Cation[114]. One Na^{+} and two Mg^{2+} cations are parameterised in standard 12-6 LJ Nonbonded Model[115].

Simulation

The system building was completed with Ambertool tLEaP: Additional Na^{+} ions were introduced to counter the overall negative charge of the intasome. Finally the system was placed in a cubic simulation box filled with TIP3BOX water residues. The periodic boundary of the system was set to minimum distance of 14 Å from the intasome surface to prevent artifact interaction of the system with itself across periodic boundary. The total simulation cell size amounted to cubic cell with 143.027 Å of length in each dimension.

Simulations were run using Namd[116] software with Amber[113] forcefield on Ranger and Kraken supercomputing clusters of Teragrid/XSEDE.

The simulation parameters include a cut-off radius of 12 Å, imposed to limit computational cost. Standard Amber scaling factor of $\frac{1}{1.2} = 0.83(3)$ was applied to non-bonded (van der Waals) interaction between 1–4 bonded atoms. Long-range electrostatic interactions were treated with regular cubic Particle Mesh Ewald summation with gridsize 144 Å, slightly larger than the periodic boundary.

All MD steps were simulated in the isothermal isobaric (NPT) ensemble maintained at temperature 300 K using a Langevin thermostat with damping coefficient γ=5 ps^{-1}, and pressure 1 bar by a Berendsen barostat with relaxation time of 0.1 ps with default water compressibility $4.57 \cdot 10^{-5}$ bar^{-1}. Each simulation is performed in discrete timestep iterations of 2 fs or $2 \cdot 10^{-15}$ s.

4.1.2. Model validation

Short single simulations were run for the protein–DNA receptor, drug ligand and complete system to validate the input data. Validation included measuring root mean square deviation (RMSD) of the protein and DNA backbone, visual inspection of the active centre with a special focus on receptor–Mg–ligand bonds, ligand conformation and Zn-HHCC complex to verify structure stability.

Intasome stability analysis

In the first step the stability of the complex consisting of protein, cations and nucleic acid without the drug, was tested in short simulation. The system's energy was minimised in 2 000 steps of steepest descent in the absence of constraints and annealed from 50 to 300 K. The system was annealed again from 50 to non-physiological 550 K over 4 000 to confirm structure stability under the chosen force field in wider range of conditions.

Finally, the system was equilibrated for 200 ps with no constraints applied. Snapshots of structure were collected at each 100^{th} step of equilibration.

Drug model stability analysis

The newly parametrised drug model was validated through a short simulation to verify the structural stability of parametrisation. The drug was minimised for 5 000 steps of steepest descent in the absence of constraints and annealed from 50 to 300 K over 25 000 steps. The equilibration covered 100 ps in 50 000 simulation steps. Snapshots of the structure were collected at each 500^{th} step of equilibration.

Stability analysis of complete inhibited intasome complex

The complete complex of protein, cations, nucleic acid and RLT was tested for stability. The system was minimised for 5 000 steps of steepest descent in the absence of constraints and annealed from 50 to 300 K over

25 000 steps. The equilibration covered 400 ps in 200 000 simulation steps. Snapshots of structure were collected at each 500^{th} step of equilibration.

Validation results

All tested systems show high stability during the annealing phase. The maximum measured RMSD for backbone atoms of 1.66 Å was observed at the last step of annealing at 550 K.

No unexpected behaviour was observed for either drug, intasome or complete complex. Tetrahedral Zn-HHCC complex remains stable. The active centre preserves its overall structure and while the drug explores its conformation space its binding interactions with the receptor and Mg ions are not broken. Although validity of drug parametrisation is hard to verify within the scope of this project, the fact that all structures remain stable despite relatively high temperature annealing is encouraging. It was therefore decided to proceed with production simulations.

4.1.3. Simulation

Two sets of simulations were run. No constraints were applied to any part of the simulated system in the first set of 20 replicas numbered nc10 to nc29 (hence the label **nc** for **n**o **c**onstraints). In the second set of 10 replicas numbered rd30 to rd39 movement of parts of the DNA was restricted by applying harmonic restraining potential (**rd** standing for **r**estricted **D**NA).

Minimisation and Annealing

Each simulation replica begins with 5 000 steps steepest descent phase and annealing from 50 to 300 K over 25 000 steps.

Equilibration and simulations

Each replica was simulated in 10 stages of MD, for 500 000 steps each, giving a total simulation time of 10 ns. Snapshots of system were collected at each $5\,000^{th}$ step of simulation giving 1000 frames for each

simulation. Data for 20 replicas (nc10–29) of unrestricted system were collected in total.

Restricting DNA movement

In the 10 replicas with restrained DNA (rd30–39) constraints were applied to a section at the far end of the DNA chain. The constraints emulate restriction of movement when the simulated strand is only a section of a longer chain. It also reduces the size of the phase-space for the system to explore. The eleven base-pairs most distant from the active centre on each of two DNA strands (22 in total) had been restricted with harmonic constraint force constant 4.0 $\mathrm{kcal \cdot mol^{-1} \cdot \AA^{-1}}$. The restricted part of DNA in relation to the whole structure is shown in Figure 4.4.

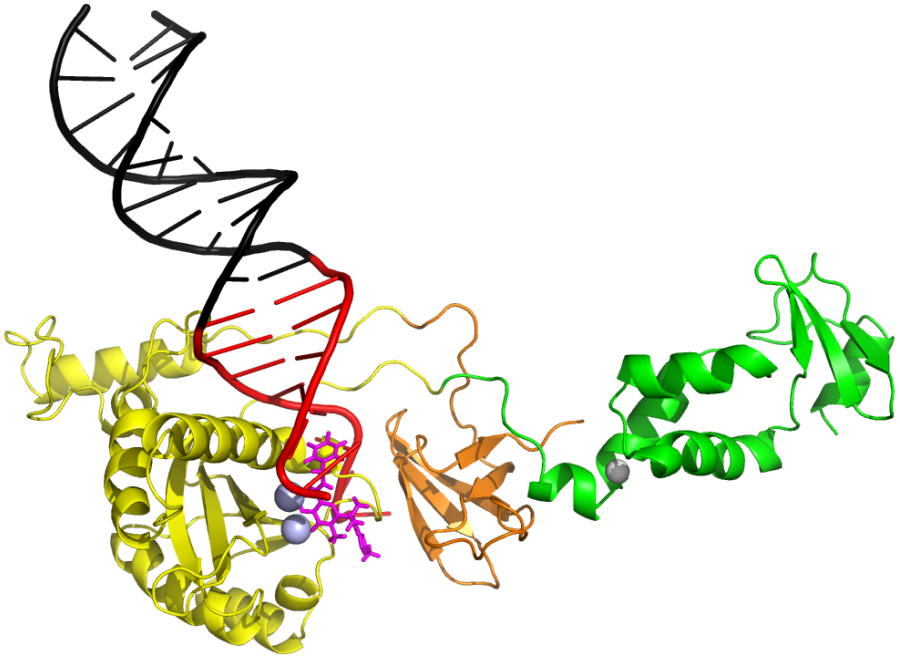

Figure 4.4. PFV structure showing restricted part of DNA in black, in relation to other elements — NTD (green), CCD (yellow), CTD (orange), DNA (red), RLT (magenta), Mg (light blue), Zn (white).

4.1.4. Binding Energy analyses

Binding free energy was calculated using the MM-PBSA method. For each simulation replica 800 trajectory snapshots collected after completed equilibration were used as an input data for analysis.

Energy contributions per residue were calculated using MM-GBSA. The trajectory snapshots analysed were the same as in the previous method.

A modification to source code of mmpbsa module of AmberTools was required to properly handle the RLT molecule. The standard version of the tool does not include parameters for Fluorine which had to be added before running analyses.

4.1.5. Principal Components Analysis

The method of principal components analysis (PCA) was applied to all *nc* trajectories in order to identify dominant factors of low convergence. The analyses were performed in ptraj. All trajectories were first concatenated and stripped of irrelevant or toublesome atoms. That includes solvent molecules (water and ions), tetrahedral Zinc, RLT and N224. Removal of N224 while not necessary for this single analysis will make it easier to compare its results with analogous analysis of N224H later. In the end analyses trajectories included 365 out of 366 amino acid residues, all DNA base pairs and two Mg ions.

The frames of concatenated trajectories of selected atoms were realigned against averaged frame and finally reduced to the first 100 principal components. In the first step only eigenvalues were compared to establish an optimal number of significant PCs.

The main analysis only takes primary PCs into account to visualise and explain the nature of movements and possibly their correlation with calculated binding energy.

A modification of ptraj source code was required to properly handle

the complete system. The standard version of the tool only works with systems no larger than 3 333 atoms. The array containing atoms positions had to be incresed from standard 10 000 and the tool had to be recompiled before running analyses.

4.2. Results

4.2.1. Minimisation and Equilibration

The RMSD of protein C_α atoms will be used as a simplified one dimensional metric of system evolution to get some insight into its behaviour, especially the equilibration in the first phase and statistical convergence in analysis phase.

The system relaxes quickly during the initial phase of simulation — as seen on Figure 4.5 — and starts levelling off after around 1 ns. Therefore, the snapshots from 200 to the end of simulation (1 000) are used as a production phase in all subsequent analyses, giving 800 snapshots or 8 ns worth data for each replica.

RMSD for some replicas (e.g. nc14) rises significantly higher (7–8 Å) than most other replicas (3–5 Å), but closer visual analysis of trajectory provides explanation in terms of NTD rotation. This motion appears to have no effects on the active site. This explanation is confirmed later by results of further analyses.

The results of simulations with restricted DNA do not differ significantly from those of unrestricted dynamics. As seen on Figure 4.6 the system equilibrates around the same time (2 ns) and stays within 2–5 Å.

4.2.2. System stability per residue

The root mean square fluctuation (RMSF) per residue averaged over whole data-sets gives us a rough idea of the relative stability of different parts of the system.

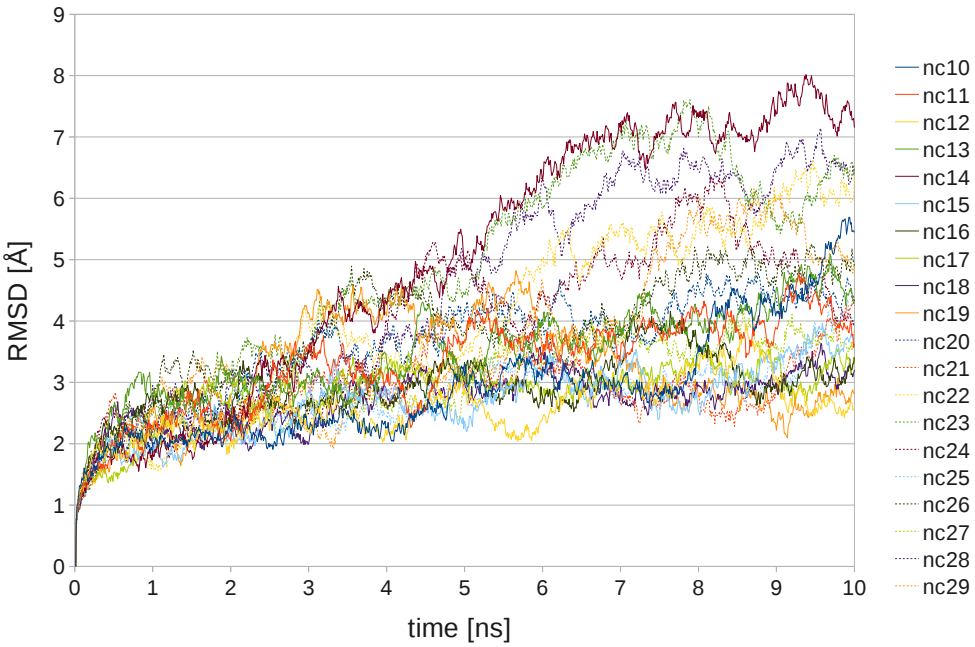

Figure 4.5. Evolution of RMSD for unrestricted PFV monomer complex C_α atoms vs first frame throughout simulation time for all 20 replicas (nc10–29). For all replicas system relaxes quickly during initial 2 ns and fluctuates freely within reasonable range afterwards.

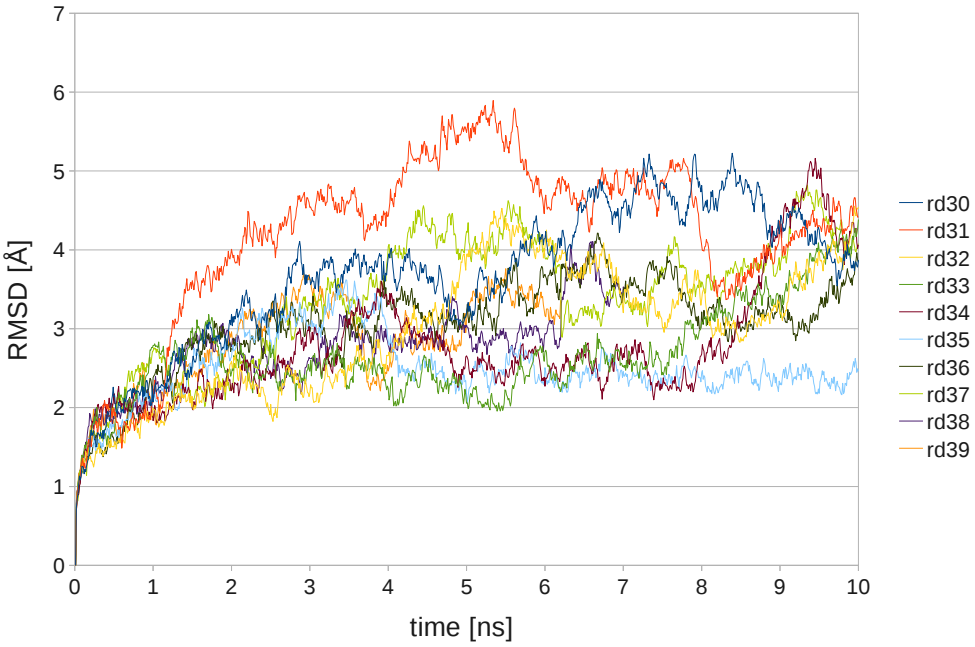

Figure 4.6. Evolution of RMSD for partially restricted PFV monomer complex C_α atoms vs first frame throughout simulation time for all 10 replicas (rd30–39). For all replicas system relaxes quickly during initial 2 ns and fluctuates freely within reasonable range afterwards.

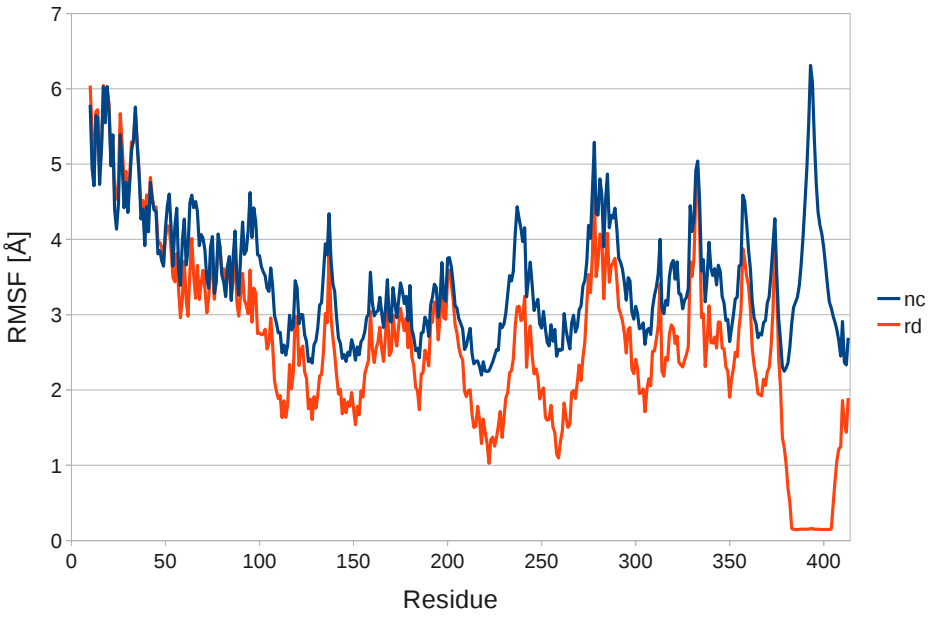

Figure 4.7. Comparison of all-atoms Root Mean Square Fluctuations per residue between unrestricted (nc) and partially restricted (rd) systems. Clearly visible difference for restricted DNA (residues 375–410). Fluctuation of protein residues within contact range of DNA is also notably reduced.

The results for both data-sets (*nc* and *rd* respectively) confirms what could be expected from the structure (Figure 4.7), exposed residues have more freedom of movement than those hidden far from the surface. Residues within the active centre — as well as those in close contact with DNA — stay within 3Å of their original positions thanks to interactions with the drug and DNA in addition to simple sterical obstacles from neighbouring residues.

Exposed residues on the other hand move as far as 4Å and farther. The NTD shows larger flexibility, as it is not stabilised by interactions with any other part of the system. Similarly, more distant DNA base-pairs fluctuate freely. The CTD, located between the domains shows intermediate behaviour.

RMSF for system with restricted DNA movement predictably differs from unrestricted system significantly. Apart from the obvious difference for restricted DNA base-pairs, and their immediate sequential neigh-

bours, protein residues within interaction distance from (not explicitly restricted parts of) the DNA, fluctuate about half as much as in the unrestricted system. The compact CCD is in direct contact with the DNA and consequently has all its residues RMSF values scaled down (residues 110–310), while the more distant NTD seems not to be affected at all, and CTD only to a limited extent. For domains position in relation to the active centre and restricted part of the DNA see Figure 4.4. All these observations are consistent with what is to be expected when constraints are applied to part of the system.

4.2.3. Binding Free Energy

The calculated binding free energies for both restrained (-21 kcal·mol^{-1}) and unrestrained (-25 kcal·mol^{-1}) definitely confirm binding character of interaction between RLT and PFV intasome. Detailed binding energy breakdown into components for both data sets is presented in Table 4.1.

system	nc		rd	
component	ΔG [kcal/mol]	stdev	ΔG [kcal/mol]	stdev
$\Delta \langle E_{vdW} \rangle$	-42.79	7.08	-42.85	7.27
$\Delta \langle E_{coulomb} \rangle$	-49.72	20.16	-40.58	9.51
$\Delta \langle E_{MM} \rangle$ subtotal	-92.51	21.70	-83.43	12.39
polar	71.99	16.39	67.01	9.22
nonpolar	-5.34	0.39	-5.43	0.42
$\Delta \langle G_{PBSA} \rangle$ solvation subtotal	66.65	16.25	61.58	9.17
total	-25.87	10.01	-21.85	9.44

Table 4.1. Binding free energy components comparison between unrestricted system (nc) and system with restricted DNA movement (rd). Detailed desctiption of components in section 2.6.

The 4 kcal·mol^{-1} of difference in total binding energy between both datasets seems significant. However, the high standard deviation of the results means they require more cautious interpretation.

More detailed analysis of gathered numerical data was performed by calculating the distribution of energy values for all collected data-points. As seen in Figure 4.8 both datasets distributions converge to normal

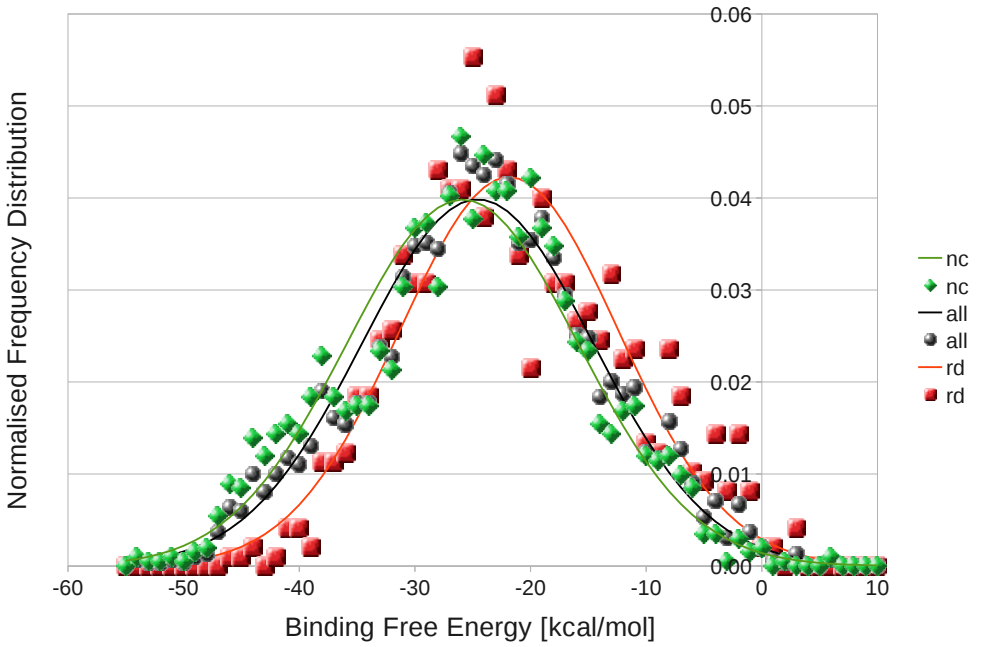

Figure 4.8. Normalised Frequency distribution function. Relatively good convergence to normal distribution may still be improved by additional data. Distribution for both unrestrained (nc) and partially restrained (rd) systems are plotted separately, as well as distribution for both datasets treated together (all).

distributions. The convergence is far from ideal though, especially the peak is heavily distorted. Additional data is required for better coverage of phase-space to improve the resolution of distribution. Most free energy values are negative (favourable) throughout all simulation replicas with exceptions constituting less then 0.5% of all data-points.

4.2.4. Distribution of results per replica

Finally, the average binding energy and standard deviation was calculated for each simulation replica separately with results presented in Figure 4.9.

Due to the limited amount of data, the difference between results of completely unrestrained and partially restrained datasets are hard to distinguish from noise. Free energy values averages for whole datasets are dwarfed by difference between single replicas within datasets.

Nevertheless, the results confirm drug binding by PFV intasome.

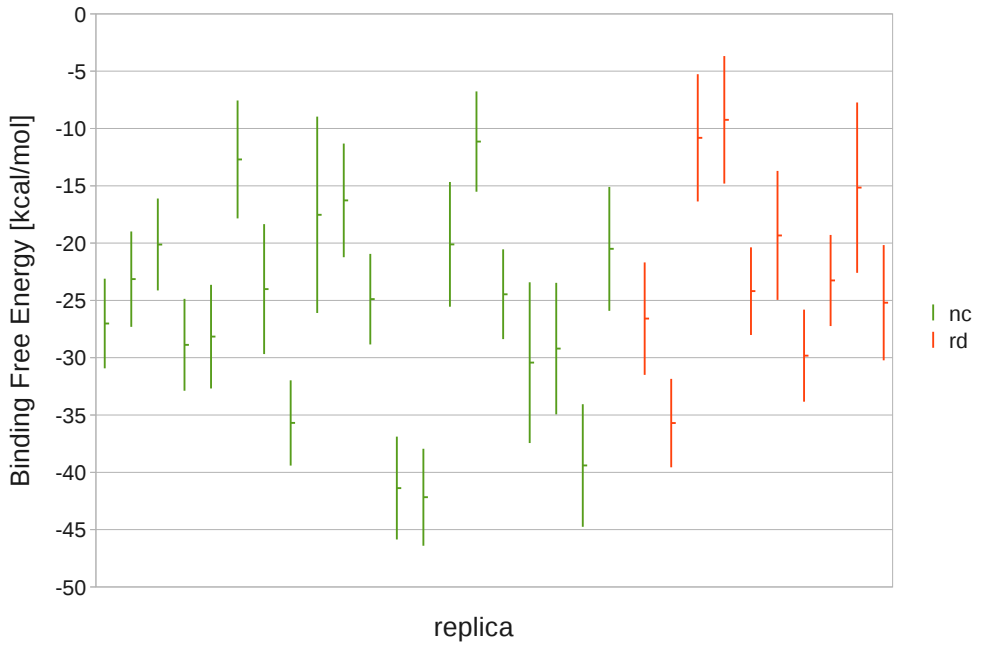

Figure 4.9. Binding Free Energy results from MM-PBSA analyses. Average binding free energy with standard deviation per replica. No significant difference between unrestricted (nc10–29) and partially restricted (rd30–39) datasets visible. Differences between single replicas averages within data-set are clearly larger than between datasets averages, big enough to explain necessity of multiple runs.

4.2.5. Per residue decomposition

The results of Binding Free Energy decomposition per residue are plotted in Figure 4.10. All values for NTD residues are expectedly null within error margin. Only eight protein residues, all of which are predictably located in CCD, and the three closest DNA base-pairs have attractive contributions larger than -0.5 kcal·mol^{-1}. Detailed numerical results are presented in Table 4.2 for protein and Table 4.3 for non-protein residues. Among those significant contributors are all three of D,D(35)E motif residues and two notable residues (G187, Y212) identified for giving resistance to HIV mutants in analogous positions. All except one are preserved between PFV and HIV-1 integrases. Positions of contributing residues are shown in Figure 4.11. All contributing residues are located within the active centre area, and are in direct contact with either RLT or Mg.

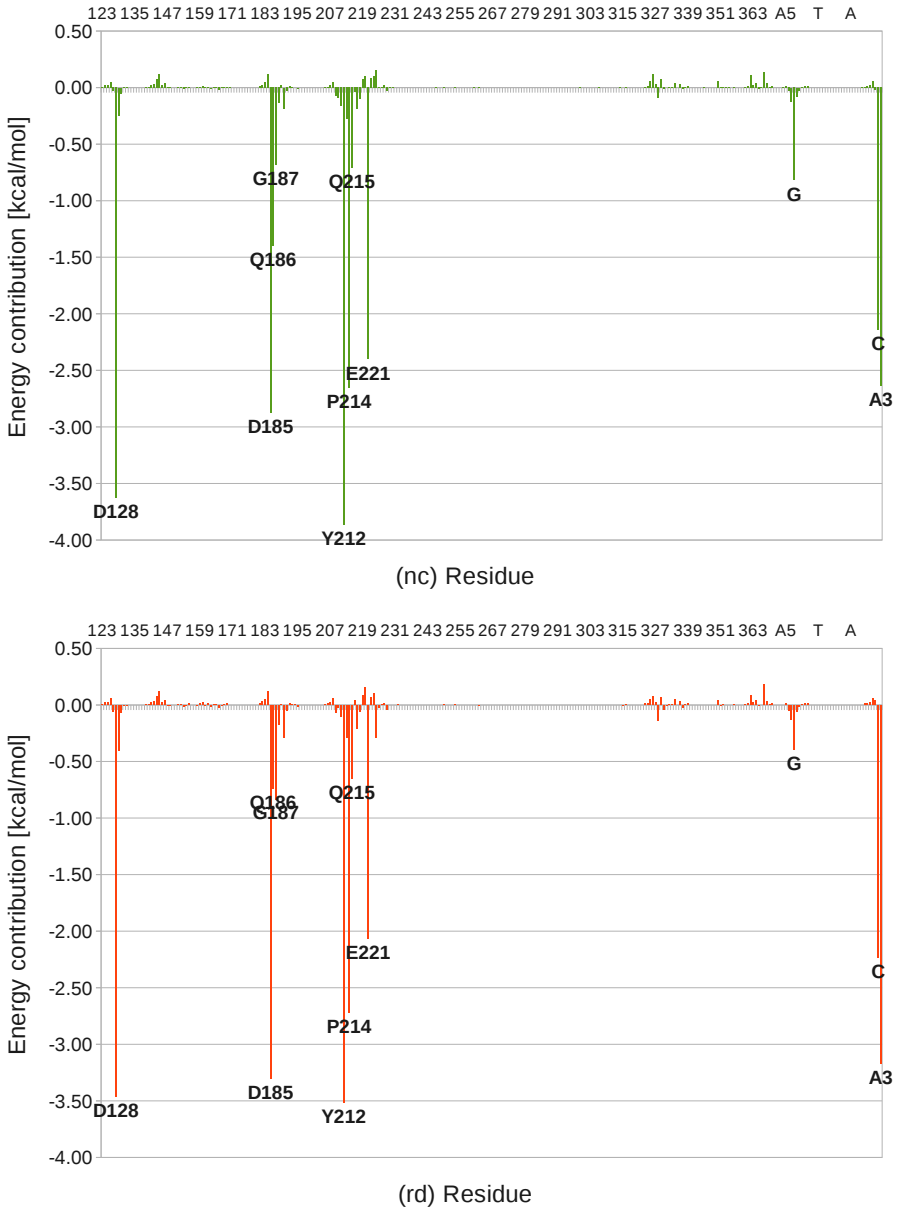

Figure 4.10. Binding free energy decomposition per residue. Negligible contribution of NTD residues not presented. All labelled residues with significant contribution values are in direct contact with either RLT or Mg.

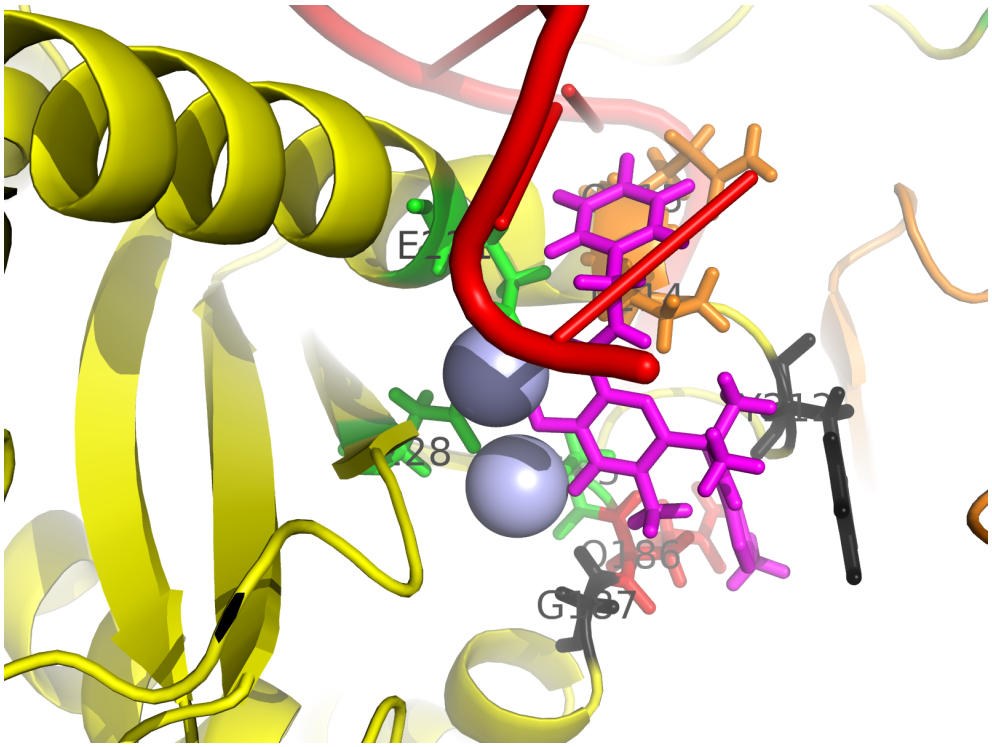

Figure 4.11. Residues contributing to calculated binding energy, in colours — DDE motif (green), residues with reported resistance giving mutations G187 and Y212 (black), non-conserved Q186 (red), P214 and Q215 (orange) in relation to other elements — CCD (yellow), DNA (red), RLT (magenta), Mg (light blue).

ΔG [kcal/mol]		protein		
nc	rd	PFV	HIV-1	note
-3.63	-3.46	D128	D64	DDE
-2.86	-3.30	D185	D116	DDE
-1.42	-0.74	Q186	N117	only non-conserved residue
-0.68	-0.83	G187	G118	MK-2048 resistant G118R [117]
-3.86	-3.51	Y212	Y143	RLT resistant Y143R[118]
-2.66	-2.72	P214	P145	
-0.72	-0.65	Q215	Q146	
-2.40	-2.07	E221	E152	DDE

Table 4.2. Residues with significant binding energy contribution and their HIV-1 IN analogues.

ΔG [kcal/mol]		
nc	rd	residue
-0.82	-0.40	DG
-2.14	-2.24	DC
-2.64	-3.17	DA3
14.73	9.15	Mg
7.10	8.33	Mg
-27.36	-25.25	RLT

Table 4.3. Binding energy contributions of non-protein residues. Residues DC and DA3 belong to CA_{OH}-3' dinucleotide. DG is a complementary base of DC.

The only unexpected result are the highly positive (unfavourable) values of ΔG for Mg^{2+} ions. It will nevertheless be ignored as the method is known for poor handling of metal ions. Removing the Mg ions from the simulation does not seem to be a good choice considering the role they play in drug/DNA binding, but ignoring the values in the final result means biased results for all other residues — all partial contributions ought to sum to the previously calculated total binding energy. However, the exact numerical values for other residues are not as important at this stage as their relative contributions.

4.2.6. Principal Components Analysis

Comparison of Principal Components eigenvalues shows a clear cutoff after third component (Figure 4.12).

Projecting original structure along eigenvectors of principal compon-

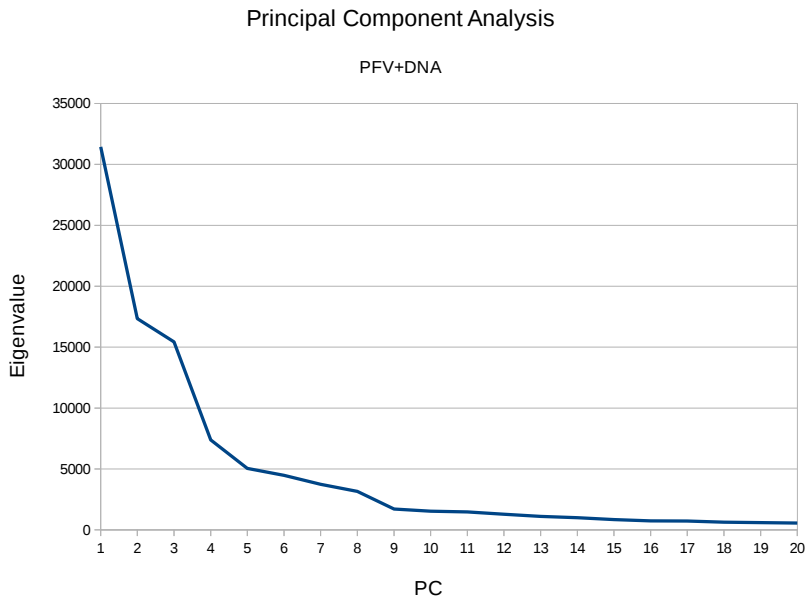

Figure 4.12. First 20 PC Eigenvalues of movement within PFV+DNA complex, Predominance of three primary eigenvectors clearly visible.

ents reveals that three primary components consist of NTD rotation around three orthogonal axis. The fourth PC includes some movement of DNA.

In conclusion, the PCA confirms the earlier results suggesting the internal movements of the complex is dominated by a NTD and DNA flexibility, which are rather the consequences of limiting the system to a monomer with freely rotating NTD and DNA outer termini, which is not a biologically meaningful result.

It is not clear whether the analysed PCs have significant effect on active centre binding affinity to the ligand. The analysis of correlation between the five principal components and binding energy components does not show any obvious connection (Table 4.4).

4.2.7. Conclusions

Initial findings confirm drug binding by the PFV intasome but low convergence means a much larger number of simulations need to be run

Energy component	PC1	PC2	PC3	PC4	PC5
$\Delta\langle E_{vdW}\rangle$	0.07	-0.03	0.16	0.04	0.05
$\Delta\langle E_{coulomb}\rangle$	0.14	-0.32	0.10	-0.29	0.13
$\Delta\langle E_{MM}\rangle$ subtotal	0.15	-0.30	0.14	-0.25	0.13
polar	-0.10	0.30	-0.16	0.15	-0.10
nonpolar	0.12	-0.01	0.09	-0.04	0.28
$\Delta\langle G_{PBSA}\rangle$ solvation subtotal	-0.10	0.30	-0.16	0.15	-0.09
ΔG total	0.16	-0.17	0.05	-0.30	0.14

Table 4.4. Correllation between principal components of *nc* trajectories movements and MM-PBSA calculated binding energy in wild type intasome monomer. PC4 gets the highest correlation coefficient for total binding energy

in parallel to obtain precise results for binding energy values. DNA flexibility seems to play a role in binding but this role is difficult to explain without additional analyses. Destabilising flexibility of DNA and NTD needs to be addressed through either restricting it by simulating stabilising interactions of larger biologically relevant complex, or through removing them from the picture completely by simulating the active centre's immediate neighbourhood only.

5. Optimal system size for accurate simulation of PFV Integrase active centre

In chapter 4 the possibility of running binding energy calculations of the PFV complex with RLT was established, but simulation of the monomer produced poorly converged results. It is essential to find what optimal subset of natural biological system could be simulated efficiently without compromising the result accuracy beyond an acceptable level. The DNA Integration process by definition involves at least two DNA strands and the enzyme itself. HIV Integrase is suggested to work in a complex of two or even four protein molecules with a number of inorganic complements (section 3.3). Atomic simulations of such large systems are very expensive in terms of both time and computing resources, thus it is important to establish a minimal subsystem sufficient for useful analyses.

The series of analyses presented in this chapter were performed on growing subsystems of the PFV Intasome dimer to establish their potential in describing HIV drug binding mechanics. The purpose of the experiment was to evaluate the ease of equilibration and the convergence of several levels of system completeness in unbiased MD simulation. Not only is it expected for more complete representation of a biological system to give more accurate results, but more importantly, because of the PFV monomer shape to observe the positive stabilising effect of complementing structures between units resulting in better convergence.

The results of unconstrained wild type monomer (**nc**) from chapter 4

were compared with the dimers built of DNA bound to CCD only (**cd**), CCD with CTD (**cc**), and with the complete dimeric system (**dr**).

5.1. Methods

The analyses described in this chapter involve both the data obtained in previous experiments and the results of simulations of modified systems.

Unless stated otherwise, all methods and parameters of MD simulations and free energy calculations in this chapter are the same as those of replicas with unconstrained DNA from chapter 4.

5.1.1. Input data and Simulation parameters

All simulation systems were based on a crystal structure of PFV IN dimer[96] (PDB:3L2T). Each protein molecule is bound with DNA double helix and a drug molecule in its active centre. Although the complete biological system is likely tetrameric, no attempt to simulate it was made due to prohibitive computational cost. Dimeric structure seems to already address the most critical shortcoming of the monomer i.e. free movement of NTD and DNA by stabilising them through the interaction between monomeric units.

Amber standard ff03.r1 forcefield was used to parametrise the dimer built of two units, consisting of a protein molecule, double DNA helix, one Na^+, and two Mg^{2+} cations each. The Zn^{2+} ion chelated by the tetrahedral HHCC motif within the NTD was parametrised as Tetrahedron-Shaped Zinc Divalent Cation[114].

The systems building was completed with Ambertool tLEaP: Additional Na^+ ions were introduced to counter the overall negative charge of the intasome. Finally the systems were placed in simulation box filled with TIP3BOX water residues to ensure a safe minimum distance from the intasome surface to the *periodic boundary condition*. Truncated octa-

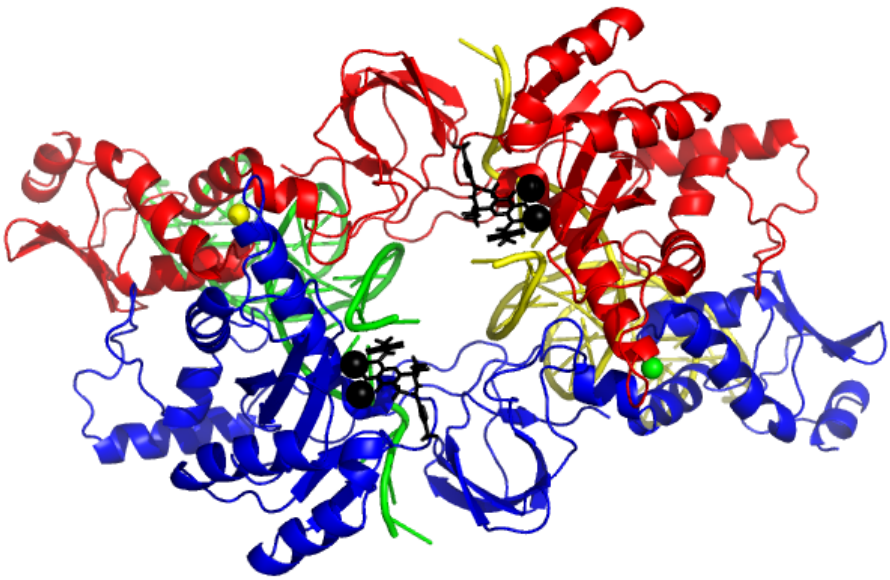

Figure 5.1. Complete structure of PFV intasome dimer with inhibitor. Parts marked by colours — Unit A: protein blue, DNA and Zn green; Unit B: protein red, DNA and Zn yellow; both MG and RLT black.

hedron was chosen as the periodic unit for complete dimer system because the system was getting too big for simulations. The truncated octahedron shape being closer to spherical than that of a cube contains less space filled with water, even with a larger minimum radius. Water molecules are not interesting in the scope of this research, yet unnecessarily increase demands on computing resources. The detailed breakdown of subsystems used in simulations is presented in Table 5.1

All simulations were run using Namd[116] software with Amber[113] forcefield on Ranger and Kraken supercomputing clusters of the Teragrid/XSEDE and the Legion computing cluster of University College London's Information Services.

Unless explicitly stated otherwise each system was simulated in 11 independent replicas.

Minimisation and Annealing

Each simulation replica begins with 5 000 steps steepest descent phase and annealing from 50 to 300 K over 25 000 steps (50 ps).

structure subset	CCD	CCD+CTD	PFV	dimer PFV
code	cd	cc	nc	dr
sequence range	117–302	100–374	10–374	2x(10–374)
atoms (protein+DNA)	4 129	5 536	7 023	14 046
RLT	1	1	1	2
Zn^{2+}	-	-	1	2
Mg^{2+}	2	2	2	4
Na^{+}	28	20	16	32
simulation box	iso 14	iso 14	iso 16	oct 16
water residues	40 802	45 343	89 925	56 136
total atoms	126 618	141 640	276 874	182 606

Table 5.1. Subsystems of wildtype PFV dimer used in simulations. Code is used in naming datafiles, usually with replica number. Number of Na^{+} ions include one ion per unit from original PDB structure with the rest added to neutralise the overall charge. Simulation box entry describes minimum distance (in Å) between the intasome surface and the wall, and box shape: cubic (iso) for most structures and truncated octahedral (oct) for the complete dimer.

Equilibration and production simulations

Each replica was simulated in stages of MD, for 100 000 – 500 000 steps each, giving a total simulation time 10 ns consisting of 5 000 000 timesteps 2 fs each. Snapshots of system were collected at each $5\,000^{th}$ step of simulation giving 1000 frames for each simulation.

5.1.2. Binding Energy

The free energy of binding between RLT and PFV active site was calculated with MMPBSA method using the same tools and settings as in chapter 4. For each simulation replica 800 trajectory snapshots collected after completed equilibration were used as an input data for analysis. Similarly, binding energy decomposition was calculated using the same method as in chapter 4. Because the dimer contains two active sites, all analyses were calculated twice, once for each site, treating the other unit together with its RLT as part of a larger receptor.

5.1.3. Symmetry

Binding free energy was calculated for each active centre separately for each simulation step to investigate the possible correlation between binding sites. Pearson's correlation coefficient between binding energy values of two active centres was tested for null hypothesis. $X = G_a$ and $Y = G_b$ were put into Equation 2.11 giving Equation 5.1:

$$r(G_a, G_b) = \frac{\sum_t \left(G_a(t) - \bar{G_a}\right)\left(G_b(t) - \bar{G_b}\right)}{\sqrt{\sum_t (G_a(t) - \bar{G_a})^2 \sum_t (G_b(t) - \bar{G_b})^2}} \tag{5.1}$$

5.2. Results

5.2.1. Equilibration

In the first step of the analysis the distribution of RMSD was compared to roughly estimate the convergence of trajectories and relative stability of all systems. While not very telling in details it may be used as a first approximation for more advanced analyses. RMSD distributions were calculated for positions of all C_α for protein and P for DNA, not just active centre. While this may seem counterintuitive, the point of this comparison is to look at the whole system. It has already been shown in chapter 4 that large scale domain movements are disruptive for the analysis even without having any obvious effect on the active centre.

	RMSD	
system	mean	stdev
cd	2.20	0.46
cc	2.14	0.43
nc	3.27	1.19
dr	1.64	0.22

Table 5.2. RMSD distribution for all sizes of simulation system. System codes are the same as in Table 5.1. The RMSD value averaged over whole trajectory gives a simple measure of system stability. High standard deviation for RMSD may indicate incomplete equilibration or low convergence.

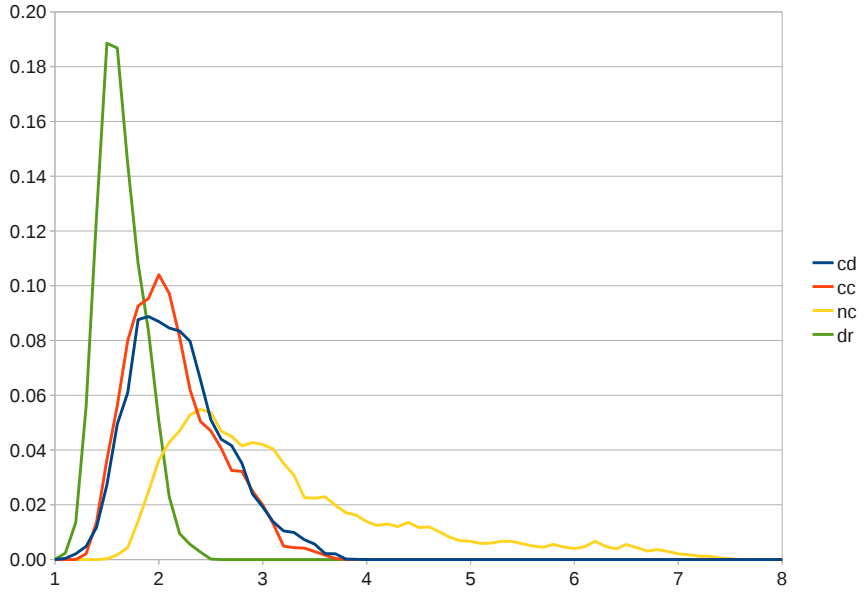

Figure 5.2. Comparison of RMSD distribution between simulated systems in relation to respective average structures. Complete dimer (*dr*) shows the highest stability and convergence of all compared systems much better than *nc* described in chapter 4. The long tail in RMSD distribution for *nc* structure needs not necessarily indicate any effect on active centre but calculation of binding energy with MMPBSA method for this system does not fare well. Incomplete dimers (*cc*, *cd*) show intermediate behaviour.

The results of comparison between systems RMS distributions are shown in Figure 5.2 and Table 5.2.

A complete monomer structure seems to have been the worst choice for free-energy calculations of those tested. Relatively high standard deviation suggests incomplete equilibration or low convergence of trajectories which result is similarly high deviation of binding energy values as found in chapter 4. While intradomain structure, including the active centre is very well preserved, the relative interdomain movements introduce a lot of variation in the atoms position measured against the average structure.

The complete dimer structure is the most stable thanks to interdomain stabilising interactions between two units, and its RMSD distribution shows a sharp and clear peak around 1.5Å. It is clearly a very good

candidate for binding energy calculation, but for a high computational price.

Similarly, incomplete structures consisting of single CCD, as well those of CCD+CTD equilibrate quickly to position with deviation within slightly over 2Å. Their structures are small and compact enough not to show interdomain movement observed with the complete monomer.

5.2.2. System stability per residue

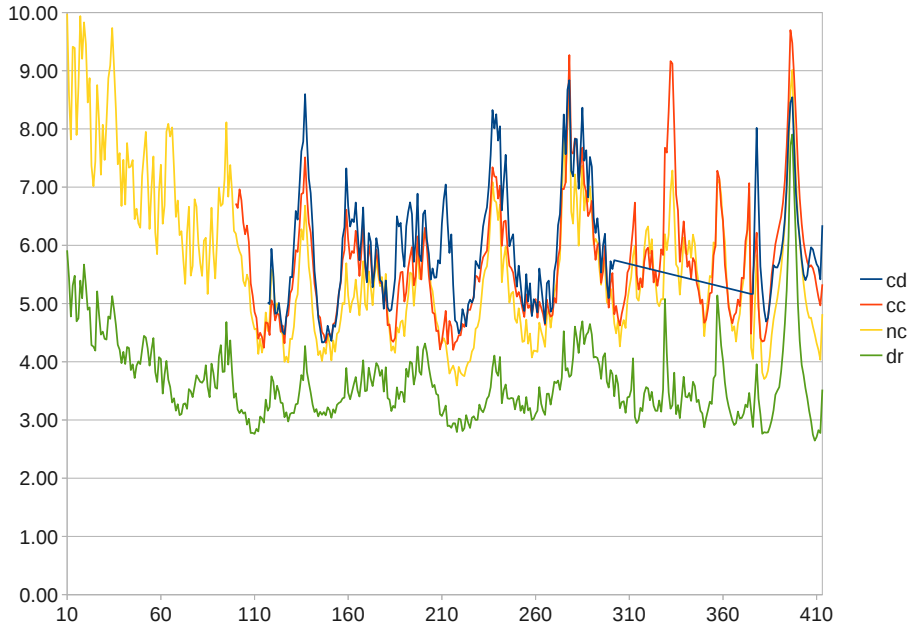

Figure 5.3. Comparison of RMSF per residue between systems of different sizes. Fluctuations of *cd* and *cc* about the same as for *nc* except for missing parts. The difference in mean deviation between complete monomer and its subsystems are caused by high fluctuation of NTD without much effect on the active site and CCD. Fluctuations of all residues within a dimer (*dr*) significantly lower than for any other system, clearly confirming stabilising effect of dimerisation.

Root Mean Square Fluctuation (RMSF) per residue averaged over each dataset presented in Figure 5.3 provides more detailed understanding of differences in stability for systems of different sizes. While RMSD showed an overall measure of internal movements for the purpose of estimating system stability, RMSF let us compare the relative fluctuations between analogous parts of the systems.

RMSF values for dimer structure is lower than monomer for all residues — not only is the system more stable as a whole but every single residue shows less fluctuation than respective residue in the monomer. The most striking difference is observed for the exposed parts of DNA or NTD — the stabilising effect of interactions between domains is clearly visible.

Intermediate overall RMSD results for *cc* and *cd* systems are now easily explained with the results of RMSF analysis. The fluctuations of atoms within CCD and CTD are not lower for these structures than for *nc*, quite the opposite in fact. For most CCD and CTD residues in *cd* and *cc* the fluctuations are actually higher than for *nc*. Removal of one or two domains from the molecule seems to weaken the rest of the structure, even though the overall RMSD value seemed to suggest the opposite. The lower RMSD values of truncated molecules are explained by very high NTD fluctuation in *nc* which obviously does not apply for these structures.

5.2.3. Binding Free Energy

The truncated sequences (*cc* and *cd*) simulations were judged to be of insufficient quality for binding energy computations, given the results of stability analysis in subsection 5.2.2. Only complete dimer (*dr*) seems to bring added value to previously analysed monomer (*nc*). Two binding sites were treated independently as different trajectories for the purpose of this analysis, doubling the total number of states taken into analysis. The results of analysis and comparison with *nc* and *rd* are presented in Table 5.3.

The overall binding energy for a dimer is consistent with the other systems. The binding energy is only 0.5 kcal/mol weaker than for *rd*. Not all components however show the same order of consistency. Both van der Waals and Coulomb components show stronger binding in the dimer and the difference is negated by a roughly equal difference for the polar solvation component. This difference is understood to be caused

system	monomer (chapter 4) nc		monomer (chapter 4) rd		dimer dr	
component	ΔG [kcal/mol]	stdev	ΔG [kcal/mol]	stdev	ΔG [kcal/mol]	stdev
$\Delta \langle E_{vdW} \rangle$	-42.8	7.0	-42.8	7.3	-44.1	4.5
$\Delta \langle E_{coulomb} \rangle$	-49.5	20.0	-40.6	9.5	-60.4	19.8
$\Delta \langle E_{MM} \rangle$ subtotal	-92.3	21.5	-83.4	12.4	-104.5	20.1
polar	71.8	16.4	67.0	9.2	88.6	17.4
nonpolar	-5.3	0.4	-5.4	0.4	-5.4	0.2
solvation subtotal	66.5	16.2	61.6	9.2	83.2	17.4
total	-25.9	9.9	-21.8	9.4	-21.3	8.5

Table 5.3. Binding free energy components comparison between monomer systems (*nc* and *rd*, see Table 4.1) and the dimer (*dr*). The dataset for the dimer is represented by sum of states of both units. The absolute values of binding energy is consistent between systems within the standard deviation margin. Stronger binding energy component is balanced by roughly equal difference in solvation energy.

by the different complex surface area. In the dimer a large area of each unit is in direct interaction with the second unit, thus decreasing the area exposed to solvent. What's important and encouraging is that the difference does not affect the overall binding energy.

5.2.4. Distribution of results per replica

Similarly to subsection 4.2.4 the distribution of results of binding energy per replica was calculated and presented in Figure 5.4. Overall distribution does not diverge from the previous experiment significantly. One notable observation is the relative independence between binding sites within a dimer. For some replicas the difference between binding sites is larger then overall variation between replicas. The correlation between binding sites will be further explored in subsequent sections.

5.2.5. Binding energy decomposition

The results of binding energy decomposition per component per active site shows an almost complete lack of contribution from the opposite unit (Table 5.4). This strongly hints at the independence of units in terms of the binding process. The result seems surprising given already

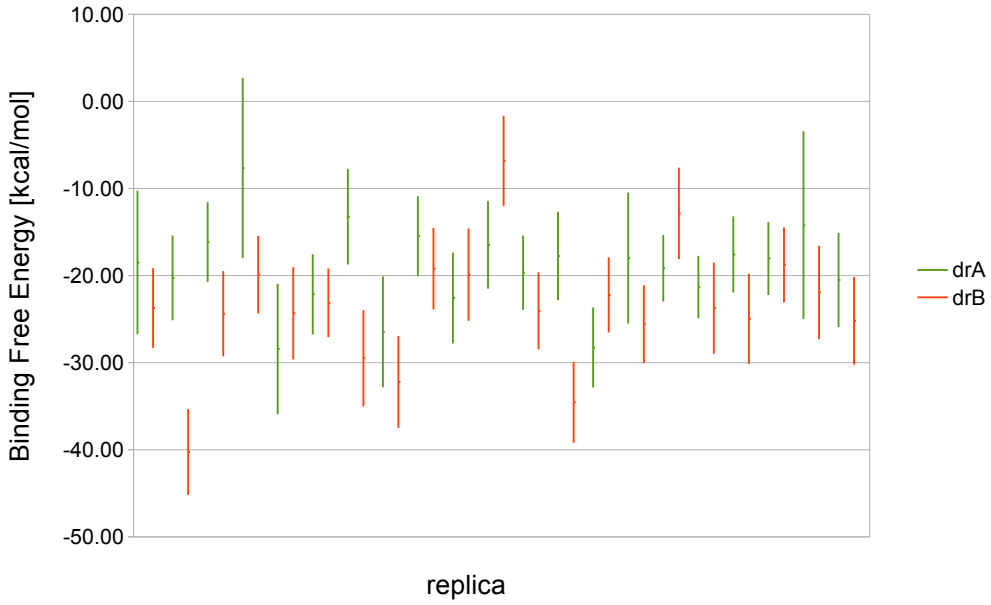

Figure 5.4. Binding Free Energy results from MM-PBSA analyses. Average binding free energy with standard deviation per replica. Differences between two binding sites for some replicas are higher then between replicas suggesting relative independence between units.

	ΔG [kcal/mol]	
subsequence	A:RLT	B:RLT
A:IN	-20.96	0.12
A:RLT	-31.16	0.01
A:DNA	-9.97	-0.05
total A	-27.92	0.11
B:IN	0.13	-20.13
B:RLT	0.01	-35.29
B:DNA	-0.02	-10.31
total B	0.16	-28.84

Table 5.4. Binding energy decomposition per unit within a dimer. Contribution of Integrase, Raltegravir and DNA of unit A and B (in rows) to binding of Raltegravir in active site A and B (in columns) respectively. Contribution of the opposite unit within a dimer seem to be completely negligible.

established importance of complete dimeric structure for the structural stability of the complex. This finding will be validated in subsequent section.

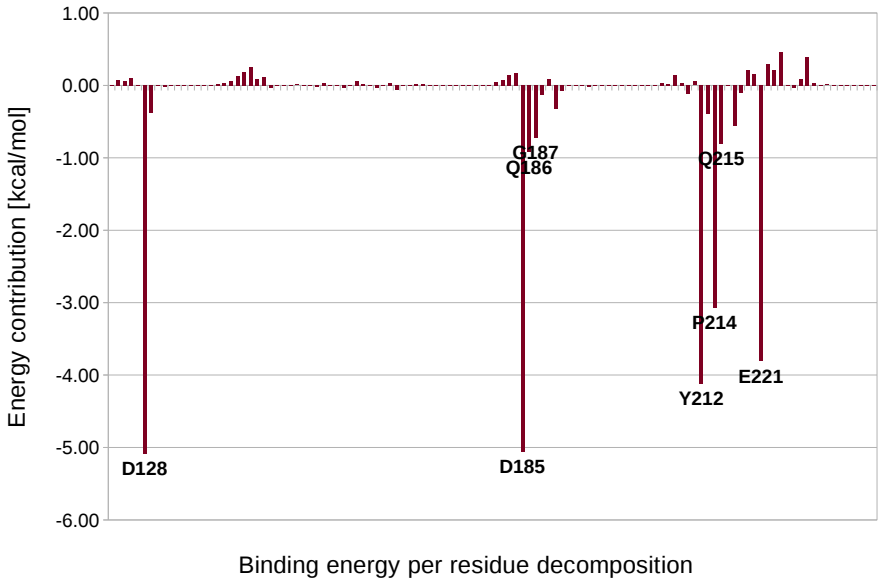

Figure 5.5. Binding free energy decomposition per residue. Only CCD residues presented. Relative contributions of labelled residues consistent with Figure 4.10.

The comparison of per residue decomposition results between monomer and dimer does not show any surprises. Slightly stronger apparent binding for some residues is offset by more repulsive Mg ions (Table 5.5). This contributions of Mg ions will be ignored as in the previous chapter. Overall, the relative contributions of protein residues presented in Figure 5.5 for the dimer are consistent with those found in subsection 4.2.5 for a monomer (Figure 4.10). The shortcoming of this method with regards to metal ions is concerning, but hopefully, limiting the focus to protein residues shall allow the identifying some effects of the mutations.

5.2.6. Symmetry

A visual analysis of trajectories show the complex breaking its symmetry easily. The calculated Pearson's correlation coefficient $r \simeq 0.09$ is marginally significant. Given almost spherical distribution (Figure 5.6)

ΔG [kcal/mol]			
nc	dr	residue	note
-3.63	-5.09	D128	DDE
-2.86	-5.06	D185	DDE
-1.42	-0.93	Q186	only non-conserved residue
-3.86	-4.12	Y212	RLT resistant Y143R[118]
-2.66	-3.07	P214	
-2.40	-3.81	E221	DDE
-17.82	-20.04	IN	whole peptide
14.73	28.62	Mg1	
7.10	6.41	Mg2	
-27.36	-33.23	RLT	
-5.74	-10.14	DNA	both strands
-29.01	-28.38	total	

Table 5.5. Comparison of binding energy decomposition per residue within a unit of a dimer with a monomer (Table 4.2, Table 4.3). Contribution of selected residues, whole peptide (IN), Magnesium ions, Raltegravir and DNA.

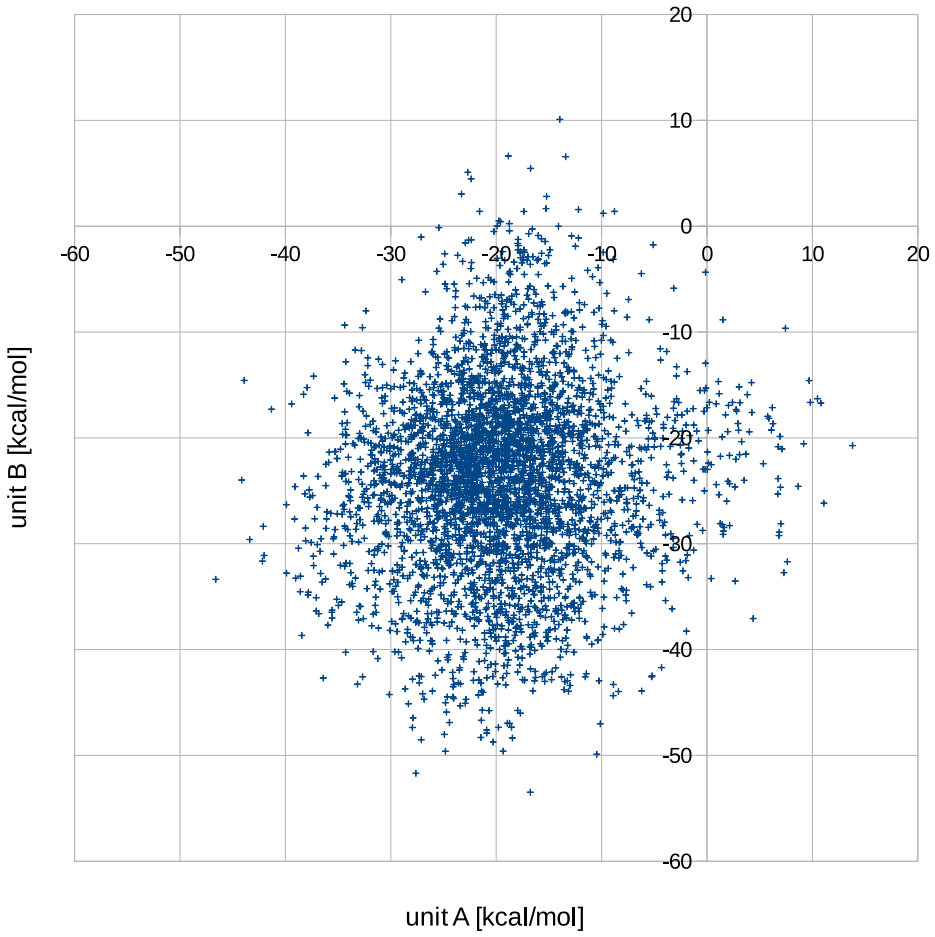

Figure 5.6. Binding Free Energy correlation between dimer units. With correlation coefficient 0.09 and almost spherical distribution on the plot the two active centres seem to be independent from each other.

the two active centres seem to be independent from each other. This confirms the results of the previous sections.

5.2.7. Conclusions

It has been shown that a single PFV intasome is a poor choice for binding energy calculations with MMPBSA analysis of MD trajectories. Dimer structure on the other hand shows much greater potential at the expense of far higher computational cost. By using an octahedral simulation box the size of the simulation system may be reduced significantly, thus enabling running more simulation replicas and getting more data using the same resources.

The implications of negligible correlation between active centres for explaining enzyme function are not clear at the moment. The direct result of this fact is however a possibility of using data for both units as independent trajectories for the purpose of some analyses, most notably MMPBSA.

At this stage of research the results' actual meaning is that they will constitute a reference for further experiments. As with most of experiments and simulations involving energy measurement, obtained values only have meaning in relation to a point of reference.

The evaluation of PFV-IN intasome dimeric structure as a viable model for HIV IN shall be approached by introducing selective mutations to explore the effects of differences between both homologues. If found feasible, the evaluation of drugs vs clinically meaningful mutations[119] will be conducted. Possible choices are Y212R — analogue of HIV-1 Y143R, for its (wildtype) binding energy contribution confirmed in subsection 4.2.5 — or N224H — analogue of N155H for its high measured experimental effect of mutation on binding energy[66].

6. Effects of HIV N155H mutant analogue PFV N224H on measured binding energy

The work described in the previous chapter was focused on establishing the experimental routine for simulating PFV Integrase complexes for the purpose of explaining resistant mutations in HIV-1 Integrase, and creating an automatic inhibitor ranking procedure. It was shown that only a simulation of the complete dimer gives a reasonable chance of obtaining credible results, given the overall complex structure and consequent stability issues with incomplete structures.

The final experimental chapter of this work describes effects of point mutation in which Asparagine in position 224 is replaced with Histidine. The analogous mutation N155H in HIV has been reported to affect drug binding in the study[66] of seven mutations (subsection 1.5.2). Only two of those —the other one being F121— are residues preserved within the lentiviral family including PFV. It is thus hoped that the molecular dynamical causes of HIV-1 resistance will have become more easily understood through this study in accordance with the main theme of this work.

The analyses include the data for wildtype proteins obtained in previous experiments as well as the results of new simulations involving mutants.

6.1. Methods

Unless stated otherwise all simulations and analyses methods and parameters in this chapter are the same as in chapter 4 and chapter 5.

6.1.1. Input data and Simulation parameters

Two experiments with mutants were conducted, monomeric system ***nh*** based on structure used in experiment *nc* and dimeric ***dh*** based on *dr* (Table 5.1). In both cases Asn residues in position 224 in respective pdb files were stripped of sidechains, renamed to His and had their sidechains reconstructed with Ambertool tLEaP.

The detailed breakdown of subsystems used in simulations is presented in Table 6.1.

	monomer		dimer	
structure subset	WT	N224H	WT	N224H
code	nc	nh	dr	dh
sequence range	10–374	10–374	2x(10–374)	2x(10–374)
atoms (protein+DNA)	7 023	7 026	14 046	14 052
RLT	1	1	2	2
Zn^{2+}	1	1	2	2
Mg^{2+}	2	2	4	4
Na^{+}	16	16	32	32
simulation box	iso 16	iso 16	oct 16	oct 16
water residues	89 925	89 923	56 136	56 135
total atoms	276 874	276 871	182 606	182 609

Table 6.1. Subsystems of PFV N224H dimer used in simulations. Code is used in naming datafiles, usually with replica number. Simulation box entry describes minimum distance (in Å) between the intasome surface and the wall, and box shape: cubic (iso) or truncated octahedral (oct). Number of Na^{+} ions include one ion per unit from original PDB structure. Wildtype systems on which mutants were built included for comparison.

All simulations were run using Namd[116] software with Amber[113] forcefield on Legion computing cluster.

Each system was simulated in 11 independent replicas.

Minimisation and Annealing

Each simulation replica begins with 5 000 steps steepest descent phase and annealing from 50 to 300 K over 25 000 steps (50 ps).

Equilibration and production simulations

Each replica was simulated in stages of MD, for 100 000–500 000 steps each, giving a total simulation time of 10 ns consisting of 5 000 000 timesteps 2 fs each. Snapshots of the system were collected at each $5\,000^{th}$ step of simulation (every 10 ps) giving 1000 frames for each simulation.

6.1.2. Binding Energy

The free energy of binding between RLT and PFV active site was calculated with MMPBSA method using the same tools and settings as in chapter 4 and chapter 5. For each simulation replica 800 trajectory snapshots collected after completed equilibration were used as an input data for analysis. Similarly, binding energy decomposition was calculated using the same method as in chapter 4 and chapter 5. Because the dimer contains two active sites, all analyses were calculated twice, once for each site, treating RLT in the other unit as part of a larger receptor.

6.1.3. Symmetry

The symmetry of binding energy values between active sites was measured as Pearson's correlation coefficient (Equation 5.1).

6.2. Results

As the simulation methods and settings were very well established at this point no initial analyses were deemed necessary. The trajectories were used to obtain binding energy immediately to compare susceptible wildtype homologue values from previous chapters against these of the molecules with RLT resistance giving mutation.

6.2.1. Binding Free Energy

system	monomer				dimer			
	nc		nh		dr		dh	
component	ΔG [kcal/mol]	std	ΔG [kcal/mol]	std	ΔG [kcal/mol]	std	ΔG [kcal/mol]	std
$\Delta \langle E_{vdW} \rangle$	-42.8	7.0	-49.2	6.8	-44.1	4.5	-43.7	5.7
$\Delta \langle E_{coulomb} \rangle$	-49.5	20.0	-52.2	30.6	-60.4	19.8	-76.4	32.9
$\Delta \langle E_{MM} \rangle$ subtotal	-92.3	21.5	-101.4	26.6	-104.5	20.1	-120.1	31.7
polar	71.8	16.4	78.5	21.5	88.6	17.4	100.8	24.6
nonpolar	-5.3	0.4	-5.6	0.2	-5.4	0.2	-5.4	0.3
solvation subtotal	66.5	16.2	72.9	21.5	83.2	17.4	95.5	24.6
total	-25.9	9.9	-28.6	9.0	-21.3	8.5	-24.7	12.7

Table 6.2. The comparison of RLT binding energy components versus both wildtype and N224H mutant in both monomer and dimer systems. RLT binds to N224H mutant stronger than to wildtype with goes opposite to expectation considering the reported resistance of N155H.

The results of drug binding turned out to be surprising. RLT shows stronger affinity to both monomer and dimer mutants than to respective wildtypes (Table 6.2). This is clearly opposite to what was expected throughout the whole project. The free energy difference between mutant and wildtype is around 3 kcal·mol^{-1} for both monomer and dimer (2.7 and 3.4 kcal·mol^{-1} respectively). As a consequence there is little point in trying to explain the mutant's resistance to RLT using these results.

6.2.2. Symmetry

The Pearson's correlation coefficient $r_{n224h} \simeq -0.12$ has higher absolute value but opposite sign to that of the wildtype ($r_{wt} \simeq 0.09$, subsection 5.2.6).

The visualisation of binding energy correlation in Figure 6.1 shows broken symmetry for one unit. The overall shape is similarly symmetrical to wildtype with the exception of the outstanding unit **B**, which skews correlation results. This finding seems to be significant because a

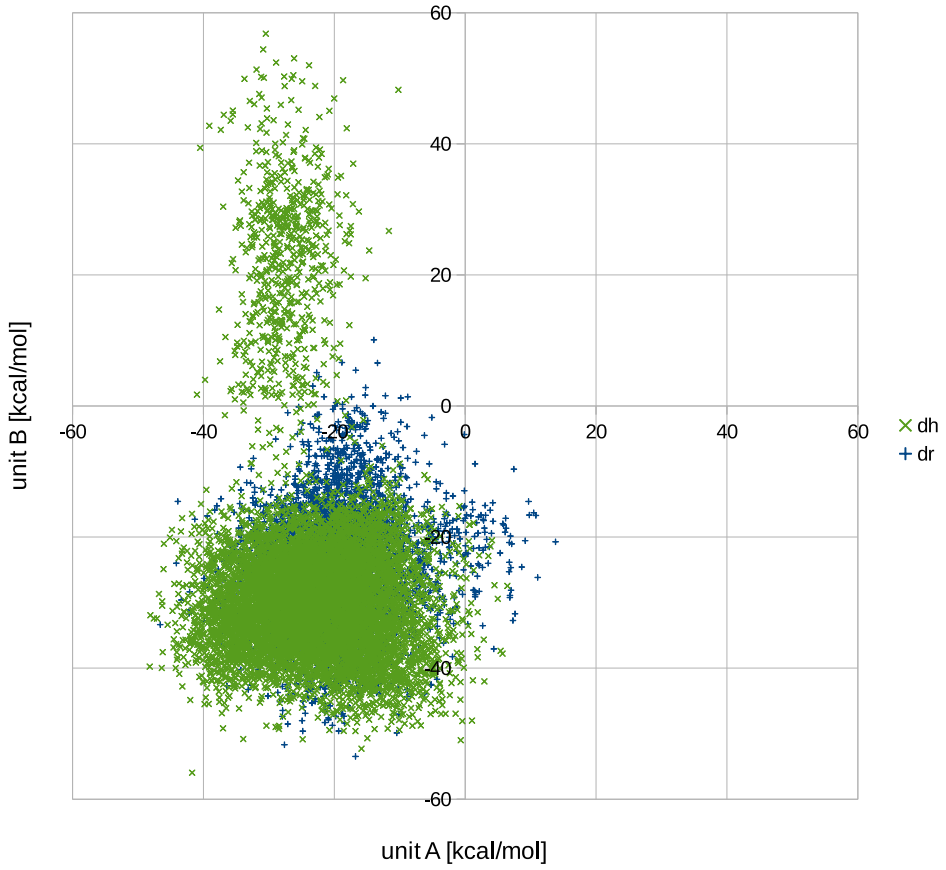

Figure 6.1. Binding Free Energy correlation between for both mutant and wild type dimer units. Clearly visible outlying single unit B of replica dh11, otherwise the shape of main dataset is similarly symmetrical for both systems.

large subset of results for the outstaning unit lies in the area of repulsive interaction.

6.2.3. Distribution of results per replica

Similarly to subsection 4.2.4 the distribution of results of binding energy per replica was calculated and presented in Figure 6.2. Nothing of significance was observed in distribution of binding energies for *nh* trajectories of the monomer, therefore given the earlier findings of its low value the monomer was abandoned at this point.

Clearly visible is the outstanding value for unit B in replica **dh11** which may signify a possible venue to follow in further research into the binding inhibition mechanism. The binding energy values averaged over

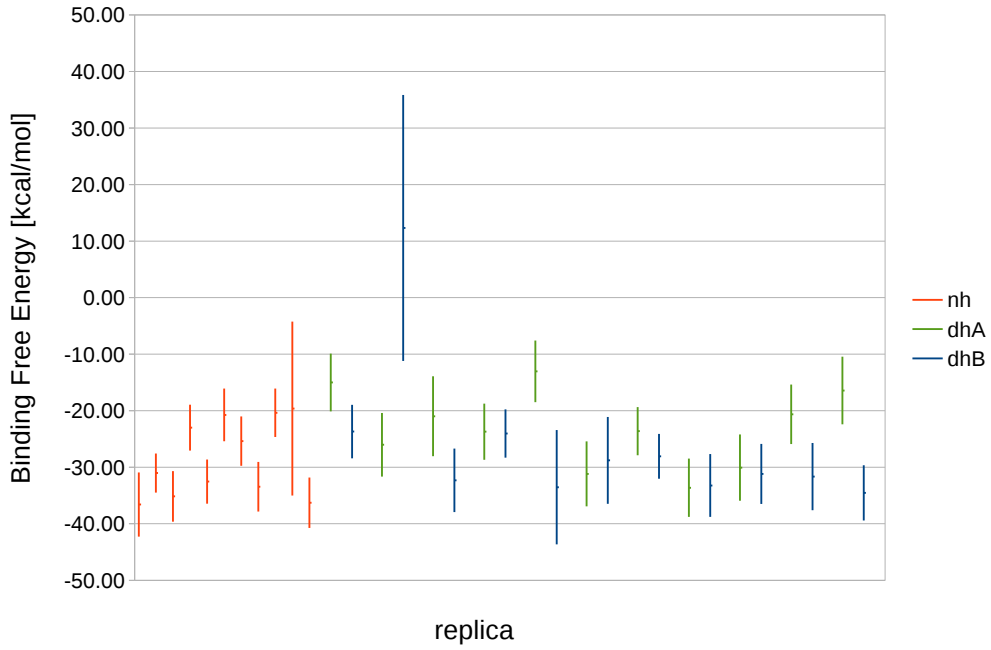

Figure 6.2. Binding Free Energy results from MM-PBSA analyses. Average binding free energy with standard deviation per replica. Clearly outstanding unit **dh11:B**. Other binding sites, including *dh11:A* not visibly affected.

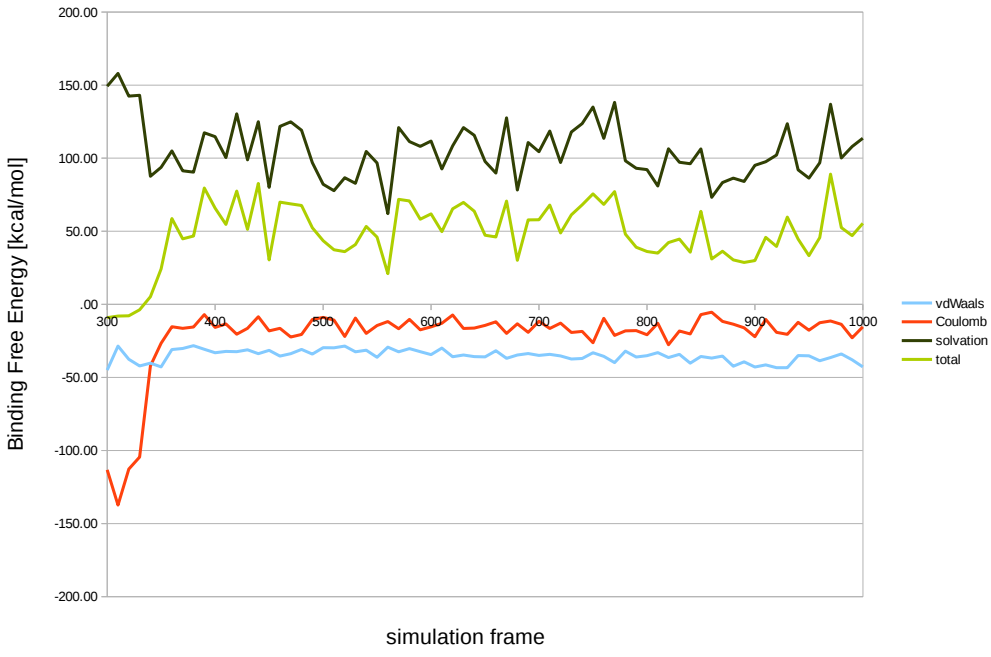

Figure 6.3. Binding Free Energy results of MM-PBSA throughout the simulation of unit *dh11:B* with breakdown into main components. Equilibration phase not included. Binding energy crosses into positive values shortly after completed equilibration and stays repulsive throughout the remainder of the simulation.

all trajectories seem to be misleading on their own without a closer look into details of results distribution. Closer analysis of MMPBSA results for simulation *dh11* shows positive binding energy in unit B throughout most of its run (Figure 6.3).

The same binding energy results with and without outstanding unit **dh11:B** are presented again in Table 6.3 and compared to wild type dimer.

system	dr		all dh		dh w/o dh11:B		dh11:B	
component	ΔG [kcal/mol]	std	ΔG [kcal/mol]	std	ΔG [kcal/mol]	std	ΔG [kcal/mol]	std
$\Delta \langle E_{vdW} \rangle$	-44.1	4.5	-43.7	5.7	-44.1	5.5	**-35.7**	4.0
$\Delta \langle E_{coulomb} \rangle$	-60.4	19.8	-76.4	32.9	-78.5	31.2	**-33.3**	38.3
$\Delta \langle E_{MM} \rangle$ subtotal	-104.5	20.1	-120.1	31.7	-122.6	29.1	**-68.9**	39.6
polar	88.6	17.4	100.8	24.6	101.5	24.4	86.3	22.7
nonpolar	-5.4	0.2	-5.4	0.3	-5.4	0.3	-5.0	0.2
solvation subtotal	83.2	17.4	95.5	24.6	96.1	24.5	81.3	22.6
total	-21.3	8.5	-24.7	12.7	-26.5	8.6	**12.3**	23.5

Table 6.3. The comparison of RLT binding energy components between all dimer systems and outstanding *dh11:B*. Unfavourable binding energy components for *dh11:B* in bold. Both VdW and Coulomb interaction in *dh11:B* bind much weaker than in all other units, including most of *dh* replicas, giving the overall repulsive effect.

While the overall binding energy for *dh11:B* is clearly unfavourable, this is only the case for one unit out of two in one of eleven replicas. The repulsive effect is not strong enough to affect the average value for a whole ensemble of simulations. In other words it would not be visible without detailed analysis of each unit separately. The clearly asymmetric behaviour of *dh11* is consistent with previously established lack of correlation between dimer binding sites.

6.2.4. Binding energy decomposition

The results of per residue decomposition for CCD are shown in Figure 6.4 together with the difference between mutant and wildtype. The

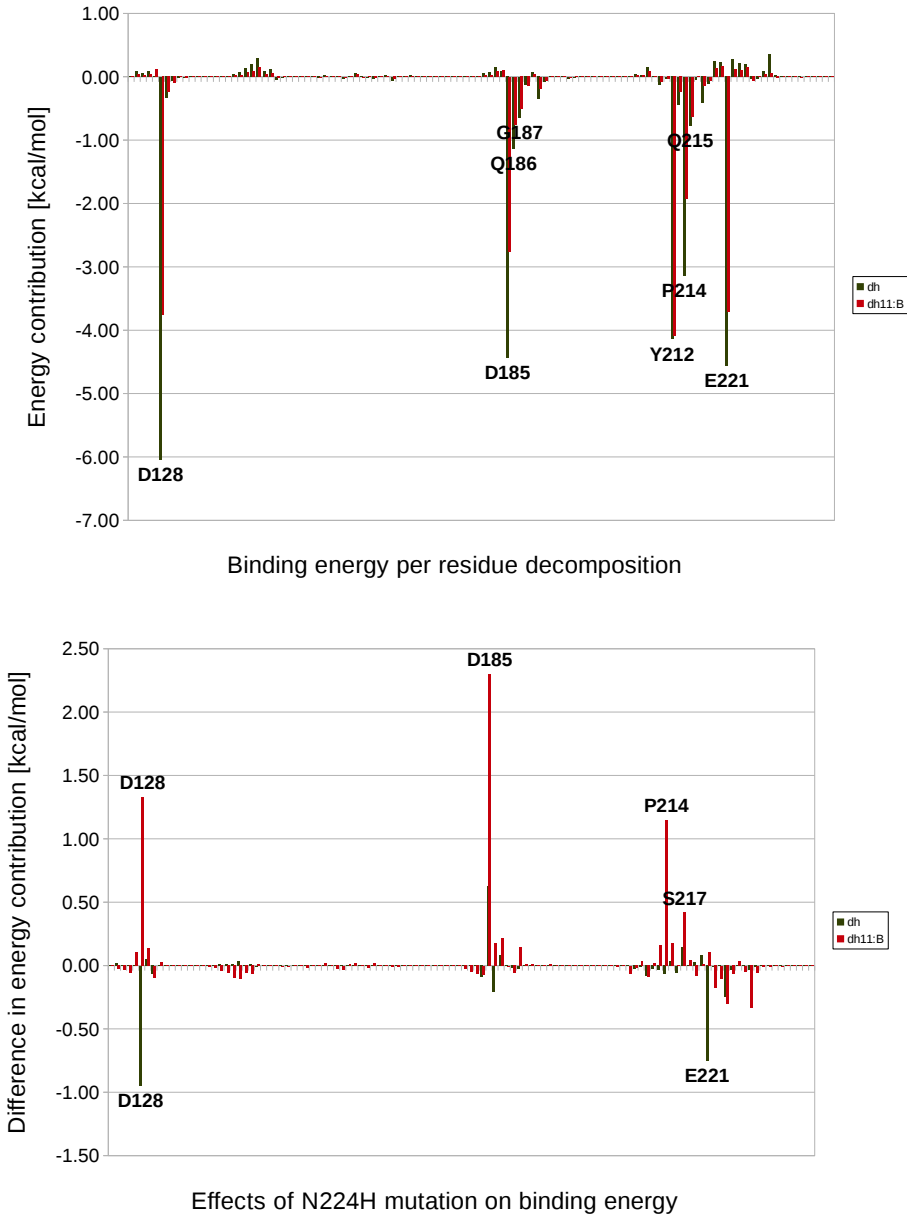

Figure 6.4. Binding free energy decomposition per residue. Only CCD residues presented. The effect of mutation, calculated as a difference between values for *dh* and *dr* is weakening of contribution of D185 even for average of all *dh* replicas. More weakening effects visible for dh11:B treated separately — D128, P214 and S217.

ΔG [kcal/mol]					
dr	dh	Δ	dh11:B	Δ	residue
-5.09	-6.04	-0.95	-3.76	1.33	D128
-5.06	-4.44	0.63	-2.76	2.30	D185
-0.93	-1.13	-0.20	-0.75	0.17	Q186
-4.12	-4.14	-0.03	-4.10	0.02	Y212
-3.07	-3.14	-0.07	-1.92	1.15	P214
-0.56	-0.42	0.14	-0.14	0.42	S217
-3.81	-4.56	-0.75	-3.71	0.10	E221
0.46	0.21	-0.25	0.15	-0.30	N224H
-1.43	-1.46	-0.02	-0.93	0.51	DG
-3.06	-3.20	-0.14	-2.03	1.03	DC
-4.03	-4.12	-0.09	-3.12	0.91	DA3
-20.04	-21.53	-1.49	-17.58	3.81	IN
28.62	32.08	3.46	22.22	-6.40	Mg1
6.41	8.36	1.95	8.75	2.35	Mg2
-33.23	-37.67	-4.44	-20.12	13.11	RLT
-10.14	-10.36	-0.23	-6.78	3.36	DNA
-28.38	-29.13	-0.75	-12.16	16.22	total

Table 6.4. Comparison of binding energy decomposition per residue within a unit of a dimer with a monomer (Table 4.2, Table 4.3). Contribution of selected residues including CA_{OH}-3', whole peptide (IN), Magnesium ions, Raltegravir and DNA. The effect of mutation is weakening of contribution of D185 even for dh average. More weakening effects visible for dh11:B — CA_{OH}-3', D128, P214 and S217.

results reveal significant weakening difference in binding contribution for several residues (Table 6.4). Surprisingly, N224H is not among them. The difference between mutants *His* and wildtype *Asn* seems to be rather strenghtening the binding. Conversely, D128 and D185 both contribute to weakening of the binding, but more notably so do P214 and S217. CA_{OH}-3' dinucleotide, together with its complementary G are also binding to mutant weaker than to wildtype. Generally, these are the results we were hoping to obtain. With the exception of the strenghtening effect of N224H itself, several residues within the active site show a weakening effect that will hopefully constitute the foundation to the explanation of viral resistance.

The results for the whole *dh* set however, are less clear. Some weakening contributions are observed for D185 and S217, but the differences are minute and completely overwhelmed by other residues.

Similarly to previous experiments very high contribution of Mg ions distorts the overall binding energy.

Spatial orientation

Residues with identified significant contribution to the difference in overall binding energy are all located within the active site Figure 6.5. They are not in direct contact with each other, however they all interact with either RLT, Mg or DNA. The nature of interaction that would result in calculated binding energies is not clear. In particular it is far from obvious how N224H mutation affects either P214 or S217 over the distance of ca 8Å (Figure 6.6).

P214 is in direct contact with RLT, which in turn interacts with N224H through either Mg or DC of CA_{OH}-3'. Interaction with S217 however, would have to be even more indirect, through either D185 or E221. Both P214 and S217 are also within 6Å from DG complementary to DC of CA_{OH}-3'.

This very interesting finding was not explored further due to time restrictions.

6.2.5. Conclusions

Given that the results for both systems are consistently opposite to predictions we are forced to admit inadequacy of the method for the given task. MMPBSA of PFV IN does not seem to be proper approach for automatic drug ranking of HIV IN inhibitors. However, many questions raised in this thesis are left unanswered, and the possibility of explaining viral resistance through MD simulations of PFV IN remains open.

Most significantly, a single replica of N224H mutant dimer shows repulsive interaction throughout most of its run. While the difference is not large enough to affect the averaged result for the purpose of automatic ranking it may yet prove to be useful in explaining the resistance of this specific mutant.

How can I possibly put a new idea into your heads,

if I do not first remove your delusions?

Robert A. Heinlein

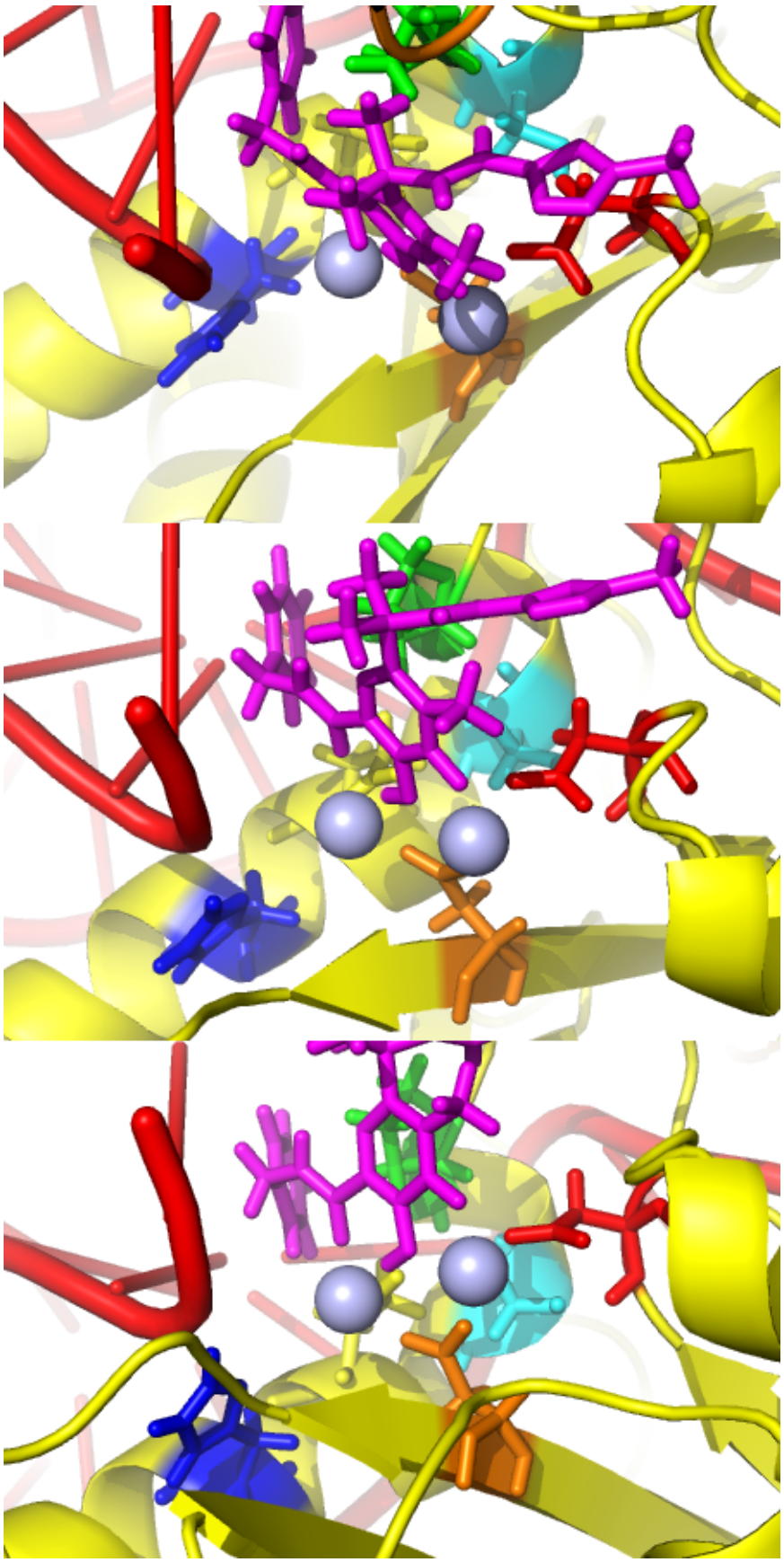

Figure 6.5. Mutant active site residues with significant in binding energy decomposition. Parts marked by colours — CCD (yellow), CTD (orange), DNA (red), RLT (magenta), Mg (light blue). DDE residues — D185 (red), D128 (orange), E221 (yellow). Other significant residues — P214 (green), S217 (cyan), N224H (dark blue). For the values of binding energy contributions see Figure 6.4 and Table 6.4.

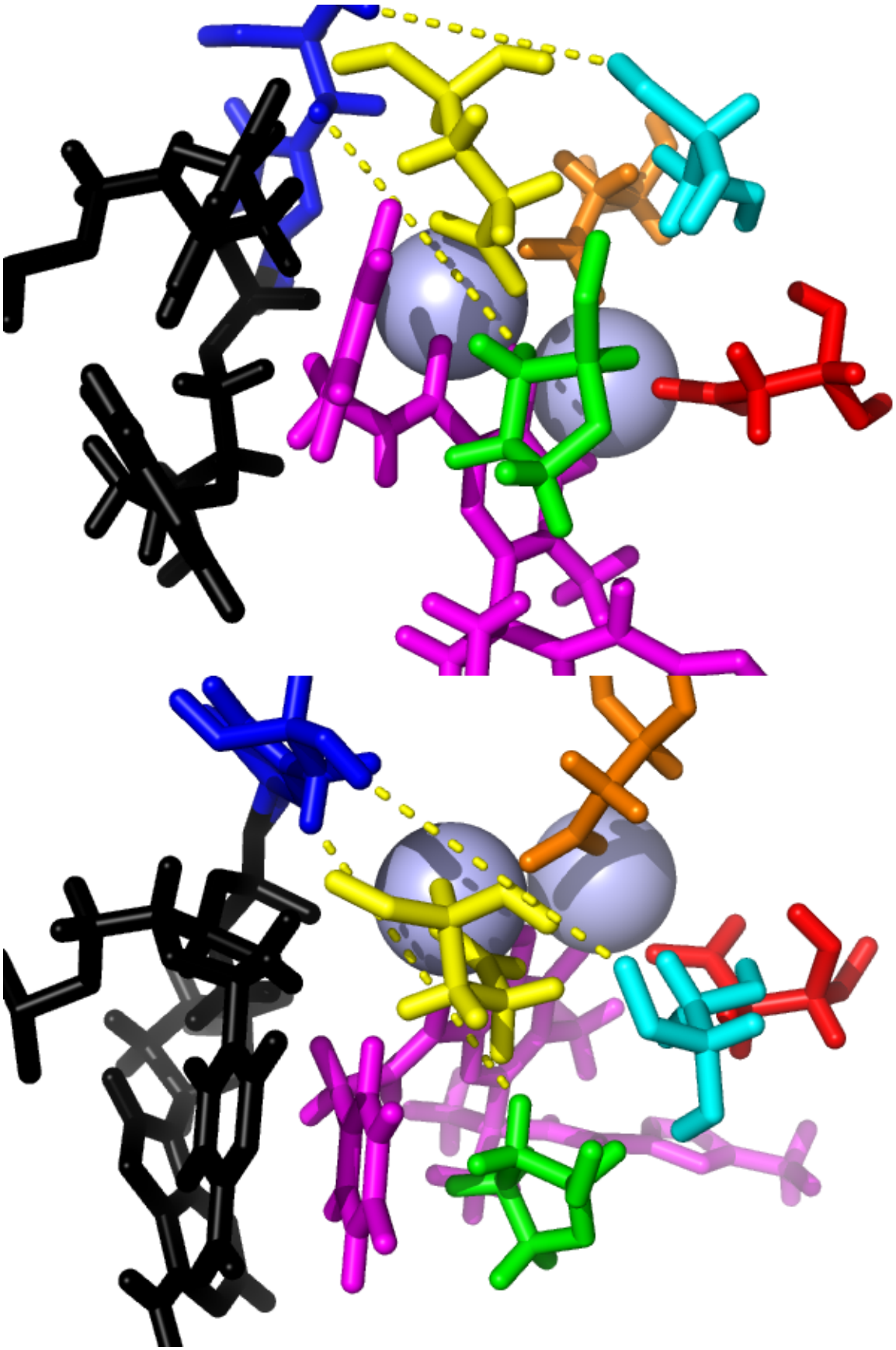

Figure 6.6. Mutant active site residues. Parts marked by colours — CA_{OH}-3' (black), RLT (magenta), Mg (light blue). Protein residues — D185 (red), D128 (orange), E221 (yellow), P214 (green), S217 (cyan), N224H (dark blue). For the values of binding energy contributions see Figure 6.4 and Table 6.4.

Bibliography

[1] G. E. P. Box, N. R. Draper, *Empirical model-building and response surfaces*, Wiley, **1987**.

[2] S. S. Frøland, P. Jenum, C. F. Lindboe, K. W. Wefring, P. J. Linnestad, T. Böhmer. HIV-1 infection in Norwegian family before 1970. *The Lancet* **1988**, *331*, 1344–1345.

[3] A. Nahmias, J. Weiss, X. Yao, F. Lee, R. Kodsi, M. Schanfield, T. Matthews, D. Bolognesi, D. Durack, A. Motulsky, P. Kanki, M. Essex. Evidence for human infection with an HTLV III/LAV-like virus in Central Africa, 1959. *The Lancet* **1986**, *327*, 1279–1280.

[4] D. Durack. Opportunistic infections and Kaposi's sarcoma in homosexual men. *New England Journal of Medicine* **1981**, *305*, 1465–1467.

[5] M. Marmor, L. Laubenstein, D. William, A. Friedman-Kien, R. D. Byrum, S. D'Onofrio, N. Dubin. Risk factors for Kaposi's sarcoma in homosexual men. *The Lancet* **1982**, *319*, 1083–1087.

[6] F. Barre-Sinoussi, J. Chermann, F. Rey, M. Nugeyre, S. Chamaret, J. Gruest, C. Dauguet, C. Axler-Blin, F. Vezinet-Brun, C. Rouzioux, W. Rozenbaum, L. Montagnier. Isolation of a T-lymphotropic retrovirus from a patient at risk for acquired immune deficiency syndrome (AIDS). *Science* **1983**, *220*, 868–871.

[7] R. Gallo, P. Sarin, E. Gelmann, M. Robert-Guroff, E. Richardson, V. Kalyanaraman, D. Mann, G. Sidhu, R. Stahl, S. Zolla-Pazner, J. Leibowitch, M. Popovic. Isolation of human T-cell leukemia virus in acquired immune deficiency syndrome (AIDS). *Science* **1983**, *220*, 865–867.

[8] B. M. Kuehn. UNAIDS Report: AIDS Epidemic Slowing, But Huge Challenges Remain. *The Journal of the American Medical Association* **2006**, *296*, 29–30.

[9] UNAIDS, *UNAIDS Report on the global AIDS epidemic*, Tech. Rep., Joint United Nations Programme on HIV/AIDS, **2010**.

[10] M. Worobey, P. Telfer, S. Souquière, M. Hunter, C. A. Coleman, M. J. Metzger, P. Reed, M. Makuwa, G. Hearn, S. Honarvar, P. Roques, C. Apetrei, M. Kazanji, P. A. Marx. Island Biogeography Reveals the Deep History of SIV. *Science* **2010**, *329*, 1487–1487.

[11] P. M. Sharp, E. Bailes, R. R. Chaudhuri, C. M. Rodenburg, M. O. Santiago, B. H. Hahn. The origins of acquired immune deficiency syndrome viruses: where and when? *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* **2001**, *356*, 867–876.

[12] D. Baltimore. Viral RNA-dependent DNA Polymerase: RNA-dependent DNA Polymerase in Virions of RNA Tumour Viruses. *Nature* **1970**, *226*, 1209–1211.

[13] H. M. Temin, S. Mizutani. RNA-Dependent DNA Polymerase in Virions of Rous Sarcoma Virus. *Nature* **1970**, *226*, 1211–1213.

[14] F. Crick. Central dogma of molecular biology. *Nature* **1970**, *227*, 561–563.

[15] F. Crick. On protein synthesis. *Symposia of the Society for Experimental Biology* **1958**, *12*, 138–163.

[16] S. H. Hughes, P. R. Shank, D. H. Spector, H.-J. Kung, J. M. Bishop, H. E. Varmus, P. K. Vogt, M. L. Breitman. Proviruses of avian sarcoma virus are terminally redundant, co-extensive with unintegrated linear DNA and integrated at many sites. *Cell* **1978**, *15*, 1397–1410.

[17] P. Lewis, M. Hensel, M. Emerman. Human immunodeficiency virus infection of cells arrested in the cell cycle. *EMBO Journal* **1992**, *11*, 3053–3058.

[18] L. Naldini, U. Blomer, P. Gallay, D. Ory, R. Mulligan, F. H. Gage, I. M. Verma, D. Trono. In Vivo Gene Delivery and Stable Transduction of Nondividing Cells by a Lentiviral Vector. *Science* **1996**, *272*, 263–267.

[19] R. Sikorski, R. Peters. Gene therapy:Treating with HIV. *Science* **1998**, *282*, 1438a–.

[20] R. G. Amado, I. S. Y. Chen. Biomedicine: Lentiviral Vectors–the Promise of Gene Therapy Within Reach? *Science* **1999**, *285*, 674–676.

[21] E. O. Freed. HIV-1 Gag Proteins: Diverse Functions in the Virus Life Cycle. *Virology* **1998**, *251*, 1–15.

[22] F. M. Hughson. Enveloped viruses: A common mode of membrane fusion? *Current Biology* **1997**, *7*, R565–R569.

[23] M. Lu, S. C. Blacklow, P. S. Kim. A trimeric structural domain of the HIV-1 transmembrane glycoprotein. *Nature Structural & Molecular Biology* **1995**, *2*, 1075–1082.

[24] R. Wyatt, P. D. Kwong, E. Desjardins, R. W. Sweet, J. Robinson, W. A. Hendrickson, J. G. Sodroski. The antigenic structure of the HIV gp120 envelope glycoprotein. *Nature* **1998**, *393*, 705–711.

[25] S. P. Layne, M. J. Merges, M. Dembo, J. L. Spouge, S. R. Conley, J. P. Moore, J. L. Raina, H. Renz, H. R. Gelderblom, P. L. Narat. Factors underlying spontaneous inactivation and susceptibility to neutralization of human immunodeficiency virus. *Virology* **1992**, *189*, 695–714.

[26] P. J. Bugelski, B. E. Maleeff, A. M. Klinkner, J. Ventre, T. K. Hart. Ultrastructural evidence of an interaction between Env and Gag proteins during assembly of HIV type 1. *AIDS research and human retroviruses* **1995**, *11*, 55–64.

[27] H. R. Gelderblom, E. H. Hausmann, M. Özel, G. Pauli, M. A. Koch. Fine structure of human immunodeficiency virus (HIV) and immunolocalization of structural proteins. *Virology* **1987**, *156*, 171–176.

[28] N. L. Davis, R. R. Rueckert. Properties of a Ribonucleoprotein Particle Isolated from Nonidet P-40-Treated Rous Sarcoma Virus. *Journal of Virology* **1972**, *10*, 1010–1020.

[29] E. Fleissner, E. Tress. Isolation of a Ribonucleoprotein Structure from Oncornaviruses. *Journal of Virology* **1973**, *12*, 1612–1615.

[30] S. Bour, R. Geleziunas, M. Wainberg. The human immunodeficiency virus type 1 (HIV-1) CD4 receptor and its central role in promotion of HIV-1 infection. *Microbiology and Molecular Biology Reviews* **1995**, *59*, 63–93.

[31] J. Hoxie, J. Alpers, J. Rackowski, K. Huebner, B. Haggarty, A. Cedarbaum, J. Reed. Alterations in T4 (CD4) protein and mRNA synthesis in cells infected with HIV. *Science* **1986**, *234*, 1123–1127.

[32] R. Geleziunas, S. Bour, M. Wainberg. Cell surface down-modulation of CD4 after infection by HIV-1. *FASEB Journal* **1994**, *8*, 593–600.

[33] E. A. Berger, P. M. Murphy, J. M. Farber. Chemokine receptors as HIV-1 coreceptors: Roles in Viral Entry, Tropism, and Disease. *Annual Review of Immunology* **1999**, *17*, 657–700.

[34] W. A. O'Brien, Y. Koyanagi, A. Namazie, J.-Q. Zhao, A. Diagne, K. ldler, J. A. Zack, I. S. Y. Chen. HIV-1 tropism for mononuclear phagocytes can be determined by regions of gp120 outside the CD4-binding domain. *Nature* **1990**, *348*, 69–73.

[35] M. Endres, P. Clapham, M. Marsh, M. Ahuja, J. Turner, A. McKnight, J. Thomas, B. Stoebenau-Haggarty, S. Choe, P. Vance, T. Wells, C. Power, S. Sutterwala, R. Doms, N. Landau, J. Hoxie. CD4-independent infection by HIV-2 is mediated by fusin/CXCR4. *Cell* **1996**, *87*, 745–56.

[36] H. M. Temin. Retrovirus variation and reverse transcription: abnormal strand transfers result in retrovirus genetic variation. *PNAS* **1993**, *90*, 6900–6903.

[37] L. Mansky, H. Temin. Lower in vivo mutation rate of human immunodeficiency virus type 1 than that predicted from the fidelity of purified reverse transcriptase. *Journal of Virology* **1995**, *69*, 5087–5094.

[38] E. O. Freed. HIV-1 Replication. *Somatic Cell and Molecular Genetics* **2001**, *26*, 13–33.

[39] A. I. Dayton, J. G. Sodroski, C. A. Rosen, W. C. Goh, W. A. Haseltine. The trans-activator gene of the human T cell lymphotropic virus type III is required for replication. *Cell* **1986**, *44*, 941–947.

[40] A. G. Fisher, M. B. Feinberg, S. F. Josephs, M. E. Harper, L. M. Marselle, G. Reyes, M. A. Gonda, A. Aldovini, C. Debouk, R. C. Gallo, F. Wong-Staal. The trans-activator gene of HTLV-III is essential for virus replication. *Nature* **1986**, *320*, 367–371.

[41] D. F. Purcell, M. A. Martin. Alternative splicing of human immunodeficiency virus type 1 mRNA modulates viral protein expression, replication, and infectivity. *Journal of Virology* **1993**, *67*, 6365–6378.

[42] S. Campbell, V. Vogt. Self-assembly in vitro of purified CA-NC proteins from Rous sarcoma virus and human immunodeficiency virus type 1. *Journal of Virology* **1995**, *69*, 6487–6497.

[43] O. K. Haffar, D. J. Dowbenko, P. W. Berman. Topogenic analysis of the human immunodeficiency virus type 1 envelope glycoprotein, gp160, in microsomal membranes. *The Journal of Cell Biology* **1988**, *107*, 1677–1687.

[44] P. W. Berman, W. M. Nunes, O. K. Haffar. Expression of membrane-associated and secreted variants of gp160 of human immunodeficiency virus type 1 in vitro and in continuous cell lines. *Journal of Virology* **1988**, *62*, 3135–3142.

[45] R. L. Willey, T. Klimkait, D. M. Frucht, J. S. Bonifacinot, M. A. Martin. Mutations within the human immunodeficiency virus type 1 gp160 envelope glycoprotein alter its intracellular transport and processing. *Virology* **1991**, *184*, 319–329.

[46] R. L. Willey, J. S. Bonifacino, B. J. Potts, M. A. Martin, R. D. Klausner. Biosynthesis, cleavage, and degradation of the human immunodeficiency virus 1 envelope glycoprotein gp160. *PNAS* **1988**, *85*, 9580–9584.

[47] P. L. Earl, B. Moss, R. W. Doms. Folding, interaction with GRP78-BiP, assembly, and transport of the human immunodeficiency virus type 1 envelope protein. *Journal of Virology* **1991**, *65*, 2047–2055.

[48] H. G. Göttlinger, T. Dorfman, J. G. Sodroski, W. A. Haseltine. Effect of mutations affecting the p6 gag protein on human immunodeficiency virus particle release. *PNAS* **1991**, *88*, 3195–3199.

[49] P. Bugelski, R. Kirsh, T. Hart. HIV protease inhibitors: effects on viral maturation and physiologic function in macrophages. *Journal of Leukocyte Biology* **1994**, *56*, 374–380.

[50] O. N. Witte, D. Baltimore. Relationship of retrovirus polyprotein cleav-

ages to virion maturation studied with temperature-sensitive murine leukemia virus mutants. *Journal of Virology* **1978**, *26*, 750–761.

[51] A. H. Kaplan, M. Manchester, R. Swanstrom. The activity of the protease of human immunodeficiency virus type 1 is initiated at the membrane of infected cells before the release of viral proteins and is required for release to occur with maximum efficiency. *Journal of Virology* **1994**, *68*, 6782–6786.

[52] M. Yeager, E. M. Wilson-Kubalek, S. G. Weiner, P. O. Brown, A. Rein. Supramolecular organization of immature and mature murine leukemia virus revealed by electron cryo-microscopy: Implications for retroviral assembly mechanisms. *PNAS* **1998**, *95*, 7299–7304.

[53] H. M. Temin. Homology between RNA from Rous Sarcoma Virus and DNA from Rous Sarcoma Virus-Infected Cells. *PNAS* **1964**, *52*, 323–329.

[54] L. A. Donehower, H. E. Varmus. A mutant murine leukemia virus with a single missense codon in pol is defective in a function affecting integration. *PNAS* **1984**, *81*, 6461–6465.

[55] P. Schwartzberg, J. Colicelli, S. P. Goff. Construction and analysis of deletion mutations in the pol gene of moloney murine leukemia virus: A new viral function required for productive infection. *Cell* **1984**, *37*, 1043–1052.

[56] L. Krishnan, A. Engelman. Retroviral Integrase Proteins and HIV-1 DNA Integration. *The Journal of Biological Chemistry* **2012**, *287*, 40858–40866.

[57] X. Li, L. Krishnan, P. Cherepanov, A. Engelman. Structural biology of retroviral DNA integration. *Virology* **2011**, *411*, 194–205.

[58] E. Dournon, W. Rozenbaum, C. Michon, C. Perronne, P. De Truchis, E. Bouvet, M. Levacher, S. Matheron, S. Gharakhanian, P. Girard, D. Salmon, C. Leport, M. Dazza, B. Regnier. Effects of zidovudine in 365 consecutive patients with AIDS or AIDS-related complex. *The Lancet* **1988**, *332*, 1297–1302.

[59] D. D. Richman. Susceptibility to nucleoside analogues of

zidovudine-resistant isolates of human immunodeficiency virus. *The American Journal of Medicine* **1990**, *88*, S8–S10.

[60] T. H. Evering, M. Markowitz. Raltegravir: an integrase inhibitor for HIV-1. *Expert Opin. Investig. Drugs* **2008**, *17*, 413–422.

[61] Y. Newberry, J. J. Kelsey. Mother to Child Transmission of HIV. *Journal of Pharmacy Practice* **2003**, *16*, 182–190.

[62] C. Darwin, *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life*, John Murray, **1859**.

[63] R. J. Pomerantz. Primary HIV-1 Resistance. *JAMA: The Journal of the American Medical Association* **1999**, *282*, 1177–1179.

[64] C. Hicks, R. M. Gulick. Raltegravir: The First HIV Type 1 Integrase Inhibitor. *Clinical Infectious Diseases* **2009**, *48*, 931–939.

[65] L. M. Chirch, S. Morrison, R. T. Steigbigel. Treatment of HIV infection with raltegravir. *Expert Opin. Pharmacother.* **2009**, *10*, 1203–1211.

[66] J. Marinello, C. Marchand, B. T. Mott, A. Bain, C. J. Thomas, Y. Pommier. Comparison of Raltegravir and Elvitegravir on HIV-1 Integrase Catalytic Reactions and on a Series of Drug-Resistant Integrase Mutants†. *Biochemistry* **2008**, *47*, 9345–9354.

[67] A. R. Leach, *Molecular modelling : principles and applications*, Prentice Hall, **2001**.

[68] J. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman, D. A. Case. Development and testing of a general amber force field. *J. Comput. Chem.* **2004**, *25*, 1157–1174.

[69] M. Christen, P. H. Hünenberger, D. Bakowies, R. Baron, R. Bürgi, D. P. Geerke, T. N. Heinz, M. A. Kastenholz, V. Kräutler, C. Oostenbrink, C. Peter, D. Trzesniak, W. F. van Gunsteren. The GROMOS software for biomolecular simulation: GROMOS05. *J. Comput. Chem.* **2005**, *26*, 1719–1751.

[70] B. R. Brooks, C. L. Brooks, III, A. D. Mackerell, Jr., L. Nilsson, R. J. Petrella, B. Roux, Y. Won, G. Archontis, C. Bartels, S. Boresch, A. Caflisch, L. Caves, Q. Cui, A. R. Dinner, M. Feig, S. Fischer, J. Gao,

M. Hodoscek, W. Im, K. Kuczera, T. Lazaridis, J. Ma, V. Ovchinnikov, E. Paci, R. W. Pastor, C. B. Post, J. Z. Pu, M. Schaefer, B. Tidor, R. M. Venable, H. L. Woodcock, X. Wu, W. Yang, D. M. York, M. Karplus. CHARMM: The Biomolecular Simulation Program. *Journal of Computational Chemistry* **2009**, *30*, 1545–1614.

[71] T. Hou, J. Wang, Y. Li, W. Wang. Assessing the Performance of the MM/PBSA and MM/GBSA Methods. 1. The Accuracy of Binding Free Energy Calculations Based on Molecular Dynamics Simulations. *Journal of Chemical Information and Modeling* **2011**, *51*, 69–82.

[72] P. A. Kollman, I. Massova, C. Reyes, B. Kuhn, S. Huo, L. Chong, M. Lee, T. Lee, Y. Duan, W. Wang, O. Donini, P. Cieplak, J. Srinivasan, D. A. Case, T. E. Cheatham. Calculating Structures and Free Energies of Complex Molecules: Combining Molecular Mechanics and Continuum Models. *Accounts of Chemical Research* **2000**, *33*, 889–897.

[73] H. Gohlke, C. Kiel, D. A. Case. Insights into Protein-Protein Binding by Binding Free Energy Calculation and Free Energy Decomposition for the Ras-Raf and Ras-RalGDS Complexes. *Journal of Molecular Biology* **2003**, *330*, 891–913.

[74] W. T. Astbury, F. O. Bell. Nature of the Intramolecular Fold in Alpha-Keratin and Alpha-Myosin. *Nature* **1941**, *147*, 696–699.

[75] L. Pauling, R. B. Corey, H. R. Branson. The Structure of Proteins. *PNAS* **1951**, *37*, 205–211.

[76] C. Levinthal, *How to Fold Graciously* in *Mossbauer Spectroscopy in Biological Systems: Proceedings of a meeting held at Allerton House, Monticello, Illinois.*, University of Illinois Press (1969), pp. 22–24. `http://www-miller.ch.cam.ac.uk/levinthal/levinthal.html`.

[77] *Protein Data Bank*. `http://www.rcsb.org/pdb`.

[78] J. A. Ladias. Convergence of multiple nuclear receptor signaling pathways onto the long terminal repeat of human immunodeficiency virus-1. *Journal of Biological Chemistry* **1994**, *269*, 5944–5951.

[79] C. S. Tsai, *Introduction to Computational Biochemistry*, John Wiley & Sons, **2002**.

[80] Nielsen, Carpinteiro, Fischer, Cabeda, Porto, Gabbe. Prevalence of the C282Y and H63D mutations in the HFE gene in patients with hereditary haemochromatosis and in control subjects from Northern Germany. *British Journal of Haematology* **1998**, *103*, 842–845.

[81] N. Christopher. Hemochromatosis: A Neolithic adaptation to cereal grain diets. *Medical Hypotheses* **2008**, *70*, 691–692.

[82] M. Larkin, G. Blackshields, N. Brown, R. Chenna, P. McGettigan, H. McWilliam, F. Valentin, I. Wallace, A. Wilm, R. Lopez, J. Thompson, T. Gibson, D. Higgins. Clustal W and Clustal X version 2.0. *Bioinformatics* **2007**, *23*, 2947–2948.

[83] N. Saitou, M. Nei. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution* **1987**, *4*, 406–425.

[84] C. M. Zmasek, S. R. Eddy. ATV: display and manipulation of annotated phylogenetic trees. *Bioinformatics* **2001**, *17*, 383–384.

[85] *DeepView*. `http://spdbv.vital-it.ch/`.

[86] N. Guex, M. C. Peitsch, T. Schwede. Automated comparative protein structure modeling with SWISS-MODEL and Swiss-PdbViewer: A historical perspective. *Electrophoresis* **2009**, *30*, S162–S173.

[87] *Modeller*. `http://salilab.org/modeller/`.

[88] *The PyMOL Molecular Graphics System, Version 1.1r2pre, Schrödinger, LLC.* `http://www.pymol.org/`.

[89] J. Lubkowski, Z. Dauter, F. Yang, J. Alexandratos, G. Merkel, A. M. Skalka, A. Wlodawer. Atomic Resolution Structures of the Core Domain of Avian Sarcoma Virus Integrase and Its D64N Mutant. *Biochemistry* **1999**, *38*, 13512–13522.

[90] F. D. Bushman, B. Wang. Rous sarcoma virus integrase protein: mapping functions for catalysis and substrate binding. *Journal of Virology* **1994**, *68*, 2215–2223.

[91] Z.-N. Yang, T. C. Mueser, F. D. Bushman, C. Hyde. Crystal structure of an active two-domain derivative of rous sarcoma virus integrase. *Journal of Molecular Biology* **2000**, *296*, 535–548.

[92] Z. Chen, Y. Yan, S. Munshi, Y. Li, J. Zugay-Murphy, B. Xu, M. Witmer, P. Felock, A. Wolfe, V. Sardana, E. A. Emini, D. Hazuda, L. C. Kuo. X-ray structure of simian immunodeficiency virus integrase containing the core and C-terminal domain (residues 50-293) - an initial glance of the viral DNA binding platform. *Journal of Molecular Biology* **2000**, *296*, 521 – 533.

[93] M. Thomas, L. Brady. Hiv integrase: a target for aids therapeutics. *Trends in Biotechnology* **1997**, *15*, 167–172.

[94] S. Hare, F. Di Nunzio, A. Labeja, J. Wang, A. Engelman, P. Cherepanov. Structural Basis for Functional Tetramerization of Lentiviral Integrase. *PLoS Pathogens* **2009**, *5*, e1000515–.

[95] S. Hare, M.-C. Shun, S. S. Gupta, E. Valkov, A. Engelman, P. Cherepanov. A Novel Co-Crystal Structure Affords the Design of Gain-of-Function Lentiviral Integrase Mutants in the Presence of Modified PSIP1/LEDGF/p75. *PLoS Pathogens* **2009**, *5*, e1000259–.

[96] S. Hare, S. S. Gupta, E. Valkov, A. Engelman, P. Cherepanov. Retroviral intasome assembly and inhibition of DNA strand transfer. *Nature* **2010**, *464*, 232–236.

[97] F. Dyda, A. Hickman, T. Jenkins, A. Engelman, R. Craigie, D. Davies. Crystal structure of the catalytic domain of HIV-1 integrase: similarity to other polynucleotidyl transferases. *Science* **1994**, *266*, 1981–1986.

[98] G. Bujacz, M. Jaskólski, J. Alexandratos, A. Wlodawer, G. Merkel, R. A. Katz, A. M. Skalka. High-resolution Structure of the Catalytic Domain of Avian Sarcoma Virus Integrase. *Journal of Molecular Biology* **1995**, *253*, 333–346.

[99] S. Maignan, J.-P. Guilloteau, Q. Zhou-Liu, C. Clément-Mella, V. Mikol. Crystal structures of the catalytic domain of HIV-1 integrase free and complexed with its metal cofactor: high level of similarity of the active site with other viral integrases. *Journal of Molecular Biology* **1998**, *282*, 359–368.

[100] Y. Goldgur, R. Craigie, G. H. Cohen, T. Fujiwara, T. Yoshinaga, T. Fujishita, H. Sugimoto, T. Endo, H. Murai, D. R. Davies. Structure

of the HIV-1 integrase catalytic domain complexed with an inhibitor: A platform for antiviral drug design. *PNAS* **1999**, *96*, 13040–13043.

[101] J. C.-H. Chen, J. Krucinski, L. J. W. Miercke, J. S. Finer-Moore, A. H. Tang, A. D. Leavitt, R. M. Stroud. Crystal structure of the HIV-1 integrase catalytic core and C-terminal domains: A model for viral DNA binding. *PNAS* **2000**, *97*, 8233–8238.

[102] J.-Y. Wang, H. Ling, W. Yang, R. Craigie. Structure of a two-domain fragment of HIV-1 integrase: implications for domain organization in the intact protein. *EMBO Journal* **2001**, *20*, 7333–7343.

[103] T. M. Jenkins, A. Engelman, R. Ghirlando, R. Craigie. A Soluble Active Mutant of HIV-1 Integrase. *Journal of Biological Chemistry* **1996**, *271*, 7712–7718.

[104] D. R. Davies, I. Y. Goryshin, W. S. Reznikoff, I. Rayment. Three-Dimensional Structure of the Tn5 Synaptic Complex Transposition Intermediate. *Science* **2000**, *289*, 77–85.

[105] P. M. Pryciak, H. E. Varmus. Nucleosomes, DNA-binding proteins, and DNA sequence modulate retroviral integration target site selection. *Cell* **1992**, *69*, 769–780.

[106] H. Müller, H. Varmus. DNA bending creates favored sites for retroviral integration: an explanation for preferred insertion sites in nucleosomes. *EMBO Journal* **1994**, *13*, 4704–4714.

[107] Y. C. Bor, F. D. Bushman, L. E. Orgel. In vitro integration of human immunodeficiency virus type 1 cDNA into targets containing protein-induced bends. *PNAS* **1995**, *92*, 10334–10338.

[108] T. S. Heuer, P. O. Brown. Photo-Cross-Linking Studies Suggest a Model for the Architecture of an Active Human Immunodeficiency Virus Type 1 Integrase–DNA Complex. *Biochemistry* **1998**, *37*, 6667–6678.

[109] K. Gao, S. L. Butler, F. Bushman. Human immunodeficiency virus type 1 integrase: arrangement of protein domains in active cDNA complexes. *EMBO Journal* **2001**, *20*, 3565—3576.

[110] A. A. Podtelezhnikov, K. Gao, F. D. Bushman, J. A. McCammon. Mod-

eling HIV-1 integrase complexes based on their hydrodynamic properties. *Biopolymers* **2003**, *68*, 110–120.

[111] L. D. Luca, A. Pedretti, G. Vistoli, M. L. Barreca, L. Villa, P. Monforte, A. Chimirri. Analysis of the full-length integrase-DNA complex by a modified approach for DNA docking. *Biochemical and Biophysical Research Communications* **2003**, *310*, 1083 – 1088.

[112] *ChemAxon's Marvin.* `http://www.chemaxon.com/products/calculator-plugins/property-predictors/`.

[113] W. D. Cornell, P. Cieplak, C. I. Bayly, I. R. Gould, K. M. Merz, D. M. Ferguson, D. C. Spellmeyer, T. Fox, J. W. Caldwell, P. A. Kollman. A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *Journal of the American Chemical Society* **1995**, *117*, 5179–5197.

[114] R. R. Roe, Y.-P. Pang. Zinc's Exclusive Tetrahedral Coordination Governed by Its Electronic Structure. *Biomedical and Life Sciences* **1999**, *5*, 134–140.

[115] P. Li, B. P. Roberts, D. K. Chakravorty, K. M. Merz. Rational Design of Particle Mesh Ewald Compatible Lennard-Jones Parameters for +2 Metal Cations in Explicit Solvent. *J. Chem. Theory Comput.* **2013**, *9*, 2733–2748.

[116] J. C. Phillips, R. Braun, W. Wang, J. Gumbart, E. Tajkhorshid, E. Villa, C. Chipot, R. D. Skeel, L. Kalé, K. Schulten. Scalable molecular dynamics with NAMD. *Journal of Computational Chemistry* **2005**, *26*, 1781–1802.

[117] T. Bar-Magen, R. D. Sloan, D. A. Donahue, B. D. Kuhl, A. Zabeida, H. Xu, M. Oliveira, D. J. Hazuda, M. A. Wainberg. Identification of Novel Mutations Responsible for Resistance to MK-2048, a Second-Generation HIV-1 Integrase Inhibitor. *Journal of Virology* **2010**, *84*, 9210–9216.

[118] S. Reigadas, G. Anies, B. Masquelier, C. Calmels, L. J. Stuyver, V. Parissi, H. Fleury, M.-L. Andreola. The HIV-1 Integrase Mutations

Y143C/R Are an Alternative Pathway for Resistance to Raltegravir and Impact the Enzyme Functions. *PLoS ONE* **2010**, *5*, e10311–.

[119] R. B. Ferns, S. Kirk, J. Bennett, I. Williams, S. Edwards, D. Pillay. The dynamics of appearance and disappearance of HIV-1 integrase mutations during and after withdrawal of raltegravir therapy. *AIDS* **2009**, *23*, 2159–2164.